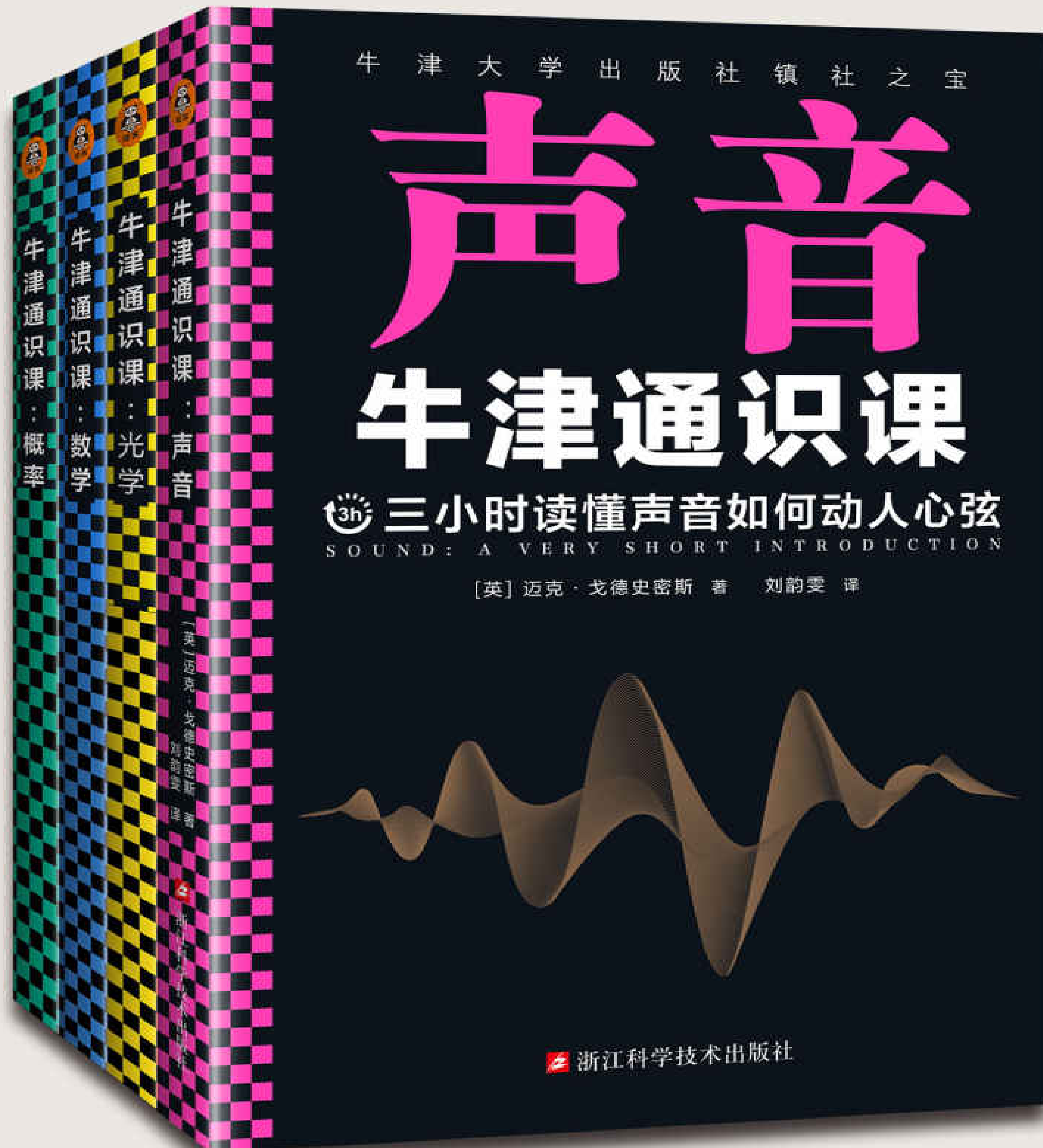
牛津通识课

（理学套装全4册）



熊猫君激发个人成长

VISIT…

LANZAROTE Caliente.com

目录

牛津通识课：声音

牛津通识课：光学

牛津通识课：数字

牛津通识课：概率

牛津通识课：声音

[英] 迈克·戈德史密斯 著
刘韵雯 译

浙江科学技术出版社

著作合同登记号 图字：11-2020-460 号

Sound: A Very Short Introduction
Copyright © Mike Goldsmith 2015
Sound: A Very Short Introduction was originally published in English in 2015. This translation is published by arrangement with Oxford University Press. Dook Media Group Limited is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

中文版权：©2021 读客文化股份有限公司
经授权，读客文化股份有限公司拥有本书的中文（简体）版权

图书在版编目（CIP）数据

牛津通识课．声音 /（英）迈克·戈德史密斯（Mike Goldsmith）著；刘韵雯译．—杭州：浙江科学技术出版社，2021.4
书名原文：Sound: A Very Short Introduction
ISBN 978-7-5341-9413-9

Ⅰ．①牛… Ⅱ．①迈… ②刘… Ⅲ．①科学知识—普及读物②声—普及读物 Ⅳ．① Z228 ② O42-49

中国版本图书馆 CIP 数据核字（2020）第 255650 号

书　　名　牛津通识课：声音
著　　者　［英］迈克·戈德史密斯
译　　者　刘韵雯

出　　版　浙江科学技术出版社
网　　址　www.zkpress.com
地　　址　杭州市体育场路 347 号
联系电话　0571-85176593
邮政编码　310006
印　　刷　北京盛通印刷股份有限公司
发　　行　读客文化股份有限公司

开　　本　787mm × 1092mm　1/32
印　　张　7
字　　数　80 000
版　　次　2021 年 4 月第 1 版
印　　次　2021 年 4 月第 1 次印刷
书　　号　ISBN 978-7-5341-9413-9
定　　价　32.00 元

责任编辑　卢晓梅　责任校对　张　宁　责任美编　金　晖　责任印务　叶文炀

图书出现倒装、缺页等印装质量问题，本社销售部负责调换

感谢

理查德·巴汉姆

格雷厄姆·卡特

马克·霍德奈特

斯蒂芬·罗宾逊

希拉里·惠特尔

声音	频率 / 赫兹
低频火山活动的声音	0.01
通过空气传导的海浪次声	0.2
大象听力范围的下限	16
人类能听到的最低频声音	20
钢琴最低音键发出的声音	27.5
有记录的最低人声	41
中央 C	261.626
钢琴的小字一组 A	440
巴赫管风琴的小字一组 A	480
有记录的最高人声	1865
乌鸫叫声的第一段	2500
钢琴最高音键发出的声音	4 186.01
人类能听到的最高频声音（80 岁）	8000
人类能听到的最高频声音（10 岁）	20 000
牙科超声清洁	25 000
一般的超声焊接	30 000
汽车的超声驻车传感器	40 000
狗的听力上限	60 000
普通蝙蝠发出的超声波	80 000
蝙蝠的听力上限	110 000
白鲸的听力上限	123 000
普通胎儿扫描	15 000 000
声学显微镜	8 000 000 000

声谱表

目录

01 声音的历史 Past Sounds
02 声音的本质 The Nature of Sound
03 和谐之声 Sounds in Harmony
04 听到声音 Hearing Sound
05 电子声 Electronic Sound
06 超声波和次声波 Ultrasound and Infrasound
07 水下和地下的声音 Sound Underwater and Underground
08 纷扰之音 Sound out of Place
术语对照
Contents
List of illustrations
The sound spectrum
Chapter 1 Past sounds
Chapter 2 The nature of sound
Chapter 3 Sounds in harmony
Chapter 4 Hearing sound
Chapter 5 Electronic sound
Chapter 6 Ultrasound and infrasound
Chapter 7 Sound underwater and underground
Chapter 8 Sound out of place
References
Further reading

01 声音的历史 Past Sounds

137亿年前的声音

声音的起源可以追溯到很久以前，远在宇宙大爆炸发生之后不久。很遗憾，宇宙大爆炸本身是寂静无声的。事实上，声波在传播媒介出现之时就形成了，那是在宇宙诞生30万年后，但直到130亿年后能够倾听到它的人类才出现。

原始声音的频率非常低，却蕴含着巨大的能量。新生宇宙的等离子体在太空中以不规则的方式开始排列，在此过程中，声音形成了。最终，在密度较大的区域诞生了星系，其中包括我们赖以生存的地球的前身，还有太阳。

从几十亿年前到地球诞生的最初数天(大约46亿年前)，有大量的声音穿过这颗固体行星的外壳和它的地下液态区，在它的大气层中反弹、转弯。最终，炎热的土地逐渐冷却，降雨汇聚成海洋，而海洋中充满了声音。最早的生物进化环境中有多种多样的声音，这些声音深刻地影响了原始生物的形态、习性和命运。

500万年前的听觉

对人类来说，听觉和触觉有明显的区别，但对海洋生物来说却并不如此。就像振动可以穿过我们的身体一样，声音能轻松地通过海洋生物的身体。鱼类依靠名为神经丘的结构感知声音，这种结构分布在它们的身体表面，除此以外，鱼类还有其他几个负责听觉的结构。神经丘含有与人耳里的毛细胞类似的细胞，鱼类依靠神经丘获得有关附

近声音强度和方向的信息。神经丘这种结构大约在5亿年前就已经演化形成。

为了探测经由空气传导的声音，鼓膜和耳蜗必不可少。因此，在大约4亿年前，当两栖动物在陆地上定居后，鼓膜和耳蜗就开始演化了。交流可能是听觉进化的主要推动力，因为声音具有远远超过视觉信号的优势。尽管一些海洋生物拥有发光和改变颜色的能力，但灯光表演远比制造噪声更具挑战性，传播范围也更小。制造噪声很容易，就像呼吸一样简单。对人类来说，呼吸发出的声音(由我们的大脑精确控制)赋予了我们说话的能力。

4万年前的音乐

对音乐的喜好是人类一种神秘且古老的乐趣。4万多年前，尼安德特人可能就已经有了笛子。在智人时期，一个没有石锣的洞穴，设施再齐全也不能算作完整的洞穴。人类的第一次歌唱甚至可能出现在语言之前。但这是为什么呢？能享受音乐并不是什么明显的演化优势。达尔文对此也感到困惑，但他提出，对音乐的品味可能是基于求偶仪式中的声音而产生的，至今也有许多人赞同这一观点。然而，还有一些人更喜欢进化心理学家史蒂芬·平克(Steven Pinker)的提议。他认为音乐是一种听觉的“芝士蛋糕”，我们享受它不是因为这种偏好能够帮助我们的祖先生存下来，而是因为许多由“芝士蛋糕”激发的情感本身就具有演化的价值。简单来说，水果的芳香显示了它已经成熟，而香醇的风味表明其中有富含能量的脂肪。人类喜欢音乐或许是因为音乐让我们想起了鸟类，鸟类的存在表明附近没有大型肉食动物。

2500年前的和声

如今，声音在我们的生活中扮演着各种各样的角色。我们的许多发明都致力于创造、传输、存储、修改或复制声音。但人类征服声音的想法不是最近才开始的。我们所知道的一些最古老的人工制品就是乐器，而且声学是最早的科学研究领域之一。公元前500年左右，毕达哥拉斯(Pythagoras)发现，当两根弦中一根的长度为另一根的一半时，拨动这两根不同长度的弦发出的声音可以和谐地融合在一起，这样的两个音之间的“距离”就是一个八度。根据定义和人们的共识，八度音程是所有音程中最和谐的。如果两根弦的长度比是其他简单的数字比例，那么其所产生的双音听起来几乎也是同样和谐的。例如，如果一根弦是另一根弦的1.5倍长，就会产生一个五度音程。

相传，在铁匠铺里，毕达哥拉斯听到铁匠们敲打铁锤时发出的和谐声音，从而发现了这一现象。作为当时一名刚崭露头角的科学家，他称量了铁匠们所用锤子的重量，发现那些发出悦耳声音的锤子的重量之间是简单的倍数关系。考虑到锤子所发出的声音的频率不是由其重量决定的，这个故事能流传至今令人惊讶。不过，无论是什么真正激起了他的兴趣，毕达哥拉斯用来研究和声的乐器其实是单弦乐器，即一种由一个可移动的琴马来控制单根弦长的乐器。

对毕达哥拉斯来说，声音的愉悦感是通过整数比来定义的，这一事实表明，数字是宇宙的关键。[1]所以他或许的确说过“万物皆数”[2]。今天的科学家会同意这一观点。而且，就毕达哥拉斯对科学方法、数学、音乐制作和声学的影响而言，他的发现可能是所有科学突破中最伟大的突破之一。

虽然任何制作或演奏弦乐器的人都知道弦的张力与长度一样能够影响弦所能发出的声音(这就是为什么要用卷弦器来为弦乐器调音)，但这一现象直到16世纪才被科学家伽利略·伽利雷(Galileo Galilei)的父亲文森佐·伽利雷(Vincenzo Galilei)量化，他发现音高随着琴

弦张力的平方根的增加而增加，现在我们知道它还取决于弦的直径和密度。

希腊人对声音的实用性很感兴趣，因为他们热衷于让别人听到自己的声音：戏剧、演说、辩论、歌曲、颂歌和宣言比比皆是。他们最伟大的符合声学结构的建筑或许是建于公元前4世纪的埃皮达鲁斯剧院。尽管该剧院的舞台中心到最后一排的距离约为60米，但在1400个座位中的任何一个地方都能清晰地听到演员的声音，要知道这些座位足足有55排之多。剧院的声学秘密就藏在座椅中。这些座椅由石灰石制成，波纹状的表面以及座椅之间的空间都有助于吸收频率低于500赫兹的声音，并反射频率更高的声音，从而降低观众窃窃私语的杂音音量，使演出效果更好。

美中不足的是，埃皮达鲁斯剧院是露天的。由于没有天花板，声音无法完全容纳于室内，所以演讲者必须非常大声，这不仅令演讲者声嘶力竭，也会剥夺声音的微妙变化。尽管传说中希腊人拥有扩音器，但这只是传说。直到17世纪70年代，扩音器才由德国的亚森纳希思·柯歇尔(Athanasius Kircher)和英国的萨缪尔·莫兰(Samuel Morland)发明出来。此外，由于剧院是露天的，外面传来的噪声也很容易对剧院内部产生干扰，不过埃皮达鲁斯剧院在演出时应该还是安静的，因为大多数当地居民都坐在剧院里了。

室内公共空间解决了上述这些问题，但又产生了新的问题，即回声和混响。要使回声成为回声，必须要使声音两次被听到之间的时间差在1/20秒以上。如果在那之前被听到，耳朵只会把两个声音当作一个更大的声音来处理。听觉中的1/20秒相当于视觉中的1/5秒，在视觉中，我们的眼睛需要间隔1/5秒才能将一个变化的东西看作两个独立的图像(因此，当相机帧的移动速度超过视觉识别独立图像的速度时，我

们就会觉得画面动起来了，这是人类的“视觉暂留现象”，即当一系列静态影像快速切换时，我们就可以看到流畅画面)。

由于空气中的声音在1/20秒内的传播距离约为10米，任何比这个距离大的房间(在任何维度上)都是可能产生回声的回音室。幸运的是，回声可以通过包有柔软织物的物体(比如观众)来减少。

当然，声音不光有科学属性和娱乐属性。即使是没有语意的声音也承载着意义，而且其中大部分意义是在史前时期就被赋予的，例如：令人感到孤独的风啸、令人震惊的痛苦尖叫、鸟儿欢快的歌声、孩子们快乐的笑声，等等。在上述这些以及其他许多情况中，演化在声音和情感之间建立了牢不可破的联系。这些情感上的依恋自古以来就被人类所利用，例如，战争中的呐喊长期以来一直存在，它既是为了让敌人不寒而栗，也是为了振奋士气和唤起战友的勇气。

现代社会：声音科学及延伸

在古希腊人对声音进行简单探索之后，直到17世纪人们才开始对声音的本质进行研究。当时，罗伯特·胡克(Robert Hooke)通过简单的论证证明了频率和音高之间是相互关联的。虽然艾萨克·牛顿(Isaac Newton)提出了一种声速方程，但这个方程是不正确的，直到1816年皮埃尔·西蒙·拉普拉斯(Pierre Simon Laplace)才推导出一个准确的版本。拉普拉斯证明声速只取决于它所经过的介质的密度和弹性(见框1)，该公式的估算准确度相当高。

框 1

紧绷的弦的振动频率为$f=\frac{1}{2l}\sqrt{\frac{T}{\mu}}$。其中，$l$是弦长，$T$是张力，$\mu$是密度。

声速$v=\frac{\sqrt{K}}{\rho}$。其中，K为弹性系数，ρ为密度。

19世纪中叶，第一个电声设备的发明引发了人类理解和控制声音的革命。麦克风、电话和扬声器相继出现，极大地刺激了研究、商业和艺术的发展。

1903年发明的二极管(第一个整流器，最初用于探测无线电信号)和1906年发明的三极管(第一个放大器)拉开了20世纪电子工程学的序幕。两次世界大战极大地促进了电子学的蓬勃发展，人类对潜艇战和舰船探测的兴趣也引发了水下声学的研究。

虽然在19世纪，人们偶尔讨论过有频率太高而人类无法听到的声音存在，但相关研究只局限于人类所能听到的最高频率的上限。尽管这些高频声音很容易由火花、哨声、喷气机或压电晶体发出，但直到第一次世界大战，当人们意识到它们可以作为声呐系统的一部分投入使用时，才对它们产生了兴趣。战争结束后不久，就有进一步的研究表明，这种声音具有一系列独特的性质，但并非所有的特性都能解释清楚，例如：它们能杀死生物，引起化学变化，产生光和热，并能使木头爆炸发出阵阵火花。它们之所以被称为超声，部分原因可能是它们拥有奇特的力量。虽然远没有X射线、α 射线和γ 射线那么有名，但

超声很快就获得了类似的声誉——借助科学工具从自然界中获得的神秘力量，

并且仍然保持着近乎超自然的神秘魅力。

直到20世纪中叶，人们才真正理解并学会利用超声波的力量。当时，电声放大技术实实在在地改变了世界。以前，受演讲者音量和演讲场地大小的限制，演讲者最多只能面对几千名观众。而现在，他们可以和成千上万公里之外的人交流——无论是即时，还是在一天或一个世纪之后。声音能够以一种前所未有的方式被捕捉、记录和分析。

声音在许多全新领域都开始崭露头角，例如一门定义不清的学科——声音艺术。它起源于20世纪初至30年代的未来主义运动，得益于电子音乐和录音技术的发展。路易吉·鲁索洛(Luigi Russolo)的《未来大协奏曲》(Gran Concerto Futuristico， 1917)是早期的一个重要例子。最近的一个例子是苏珊·菲利普斯(Susan Philipsz)的《低地》(Lowlands)，这是一首挽歌的一系列变奏曲，并在2010年赢得了特纳奖。一个相关的领域是环境音乐，通常被批评者称为**助兴音乐(muzak)**[3]，它的设计初衷是为公共场合提供合适的背景音乐。布莱恩·伊诺(Brian Eno)的专辑《环境1：机场音乐》(Ambient 1: Music for Airports， 1978)就是一个很好的例子，而超市里没完没了的甜腻圣诞颂歌以及伴随着这些音乐被迫装扮成精灵的阴郁员工则是一个坏例子。

环境音乐是人造声景的一个实例。这个概念是由穆瑞·谢弗(Murray Schafer)推广开来的。穆瑞·谢弗曾在20世纪60年代末于温哥华启动了名为“世界声景计划”的国际研究项目，这一计划在1993年促成了世界声学生态论坛的设立。谢弗具有很大的影响力，其中一部分原因在于他令人振奋的主张，即声学环境不仅可以揭示居住在其中的人的社会状况，甚至可以预测社会将如何演变。

运用谢弗的方法，经济学家、博学多才的雅克·阿塔利(Jacques Attali)提出，音乐风俗的变化预示着社会将发生更广泛的变化。这种观点由历史学家阿兰·科尔宾(Alain Corbin)进一步发展，科尔宾认为，18世纪和19世纪法国乡村的钟声塑造了那里的社会和经济关系。艺术家兼作家布兰登·拉贝尔(Brandon Labelle)说：“我的感觉是，仅仅从一种声音里就可以窥见整个历史和文化。”

更普遍的是，声景这一概念在一系列享有盛誉的学科中风行起来，虽然谢弗原本定义的术语已经在涉及相对和动态特性的声学环境应用中使用。技术历史学家艾米丽·汤普森(Emily Thompson)指出，声学环境“既是物理环境，同时也是我们感知环境的一种方式”。

在电影中，人造声景的构建在一定程度上是通过音效来实现的。自广播剧诞生以来，人造声景一直是广播剧的支柱。例如，在古希腊时期人们就能在剧院里听到人造的打雷声。在电影中，设计、制作和使声音效果与屏幕上的事件同步的工艺被称为拟音(Foley)[4]。

现在我们已经不再仅仅依赖耳朵感知声音，也不只依赖人声和机械设备来产生声音。我们可以研究和使用超低频、超高频或非常弱的声音，人们甚至都无法听到它们。我们还可以生成具有巨大能量的声束，并应用在医学、国防、地图测绘及许多其他领域。第二次世界大战以来，人们研究出了生产和引导超高强度的声束的方法，使基于声音的武器开始真正发展起来。当前仍在使用中的最著名的例子是远程声学设备(LRAD)[5]，它可以发出语言指令或令人不快的声音。在一些国家，它一直被用来对付敌人和野生动物。

随着人类社会的进步，声音得到越来越广泛的应用，噪声污染也蔓延到了世界上的许多地方。高度敏感的听觉系统曾经为我们的祖先

带来了无数便利，使我们能够享受音乐和便捷交流，但现在它也给我们带来了烦恼、压力和伤害。

因此，虽然我们已经精通声音的产生，但还远远不能控制它。为了能够控制声音，我们需要了解它的本质。

[1] 意指宇宙万物都遵循数字规则和比例。

[2] 此处原文为All is number，出自毕达哥拉斯的名言。

[3] muzak又译为穆扎克音乐，即安静柔和的背景音乐。

[4] 以音效艺术家杰克·弗利(Jack Foley)命名，一种将生活中的声音加入电影中的后期处理技术。

[5] 一种声波武器，可以驱散大规模集结的人群。

02 声音的本质 The Nature of Sound

声音的两面

森林里有一棵树倒了，但倒地声没有被人听见，这算是发出了声音吗？“声音”既是一种物理现象，又是一种感觉，这种双重意义给了刚刚的问题一个明确的答案——树倒的时候发出了、也没发出声音。声音的物理和感官方面的关系是复杂的，因为声音给我们留下的许多印象与它的物理参数有关，但又不能仅仅简化为物理参数。比如，高频音通常听起来音调更高，更大的声音听起来也更响亮。此外，从警报声到风笛声，从摇篮曲到狮子的吼叫声，许多声音对我们产生了情感上的影响，而这些影响与它们物理参数之间的关系却非常模糊。

声音在物理方面远比在情感方面更容易被人理解，所以我们将从物理学开始讲起。

压力波

声音通常是由物体的循环运动发出的，比如：扬声器的膜片不断地跳动，声带之间的间隙时而缩小时而扩大，或者吉他弦来回振动。正是这些运动向周围介质(固体、液体或气体)的传播及其在介质中的传播过程构成了声音。在某些情况下，运动的就是介质本身，比如有人在瓶口上方吹气时瓶颈里的空气就是如此。非运动源包括突然释放的热能(如爆炸或火花)和快速振荡的热源。

当扬声器的膜片运动时，膜片产生的声波具有和电信号相同的变化规律。每当膜片向外移动时，它同步挤压前面的空气分子，使它们

彼此靠近，从而形成一个高压区。这些分子接着对邻近的分子施加压力，使它们也依次靠得更近，因此一个紧密结合的分子形成脉冲(压缩波)穿过介质，接着，又因为膜片向内运动而产生一个低压区(稀疏波)。

然后膜片再次向外移动，产生第二个脉冲。膜片在1秒内由内向外移动的频率决定了声波的频率(单位是赫兹，缩写为Hz)。最简单的声波就是纯音，例如音叉发出的声音。空气压力随音叉距离变化的波形是一个正弦波，波的瞬时形状如图1所示。

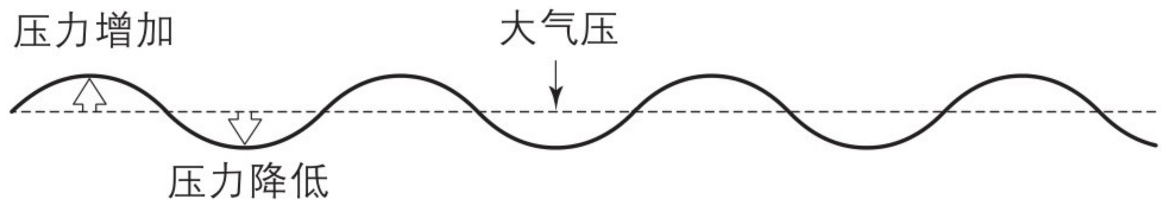


图1　声波压力示意图

相邻的两个波峰(或波谷)之间的距离定义为声波波长(λ)。声音以速度v在空气中传播，在室温下速度约为每秒340米。频率(f)由方程$f=v/\lambda$给出。空间中某一质点的压力随时间变化的曲线图也是一个正弦波，所以我们其实也可以将图1的横轴标记为“时间”。

类似图1这样的图像十分常见，而且通过这样的图像我们很容易想象出声波的某种可视图，事实上许多书就是以这种方式来使用这些图像的。然而，实际上，声波不像海浪那样有上下(横向)运动，唯一的运动是分子交替地从声源向外或向声源运动，就像牛顿摆[1]里的球一样。这种波被称为纵波，如果我们能看到空气分子的话，它们看起来就如图2所示的样子。

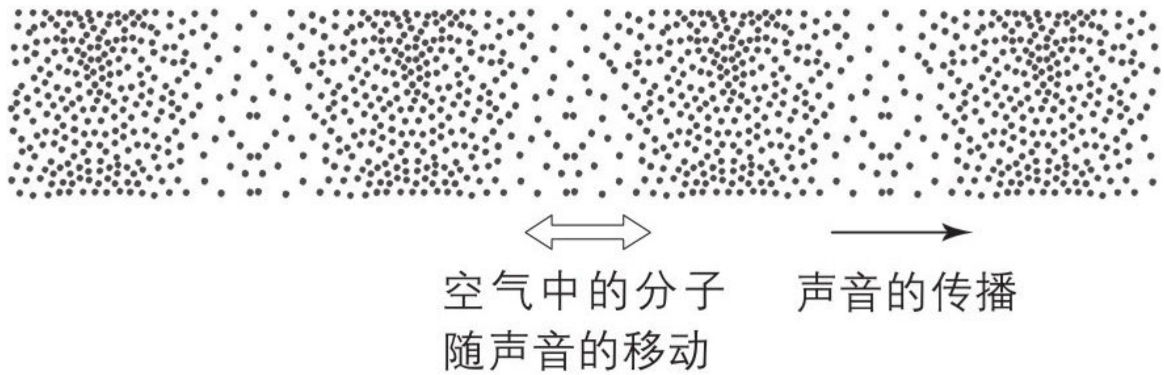


图2　从分子角度看声波

如果连续的声音起源于一个点，那么它就会像膨胀的球体一样向四面八方传播。如果探测区域很小(如麦克风隔膜或鼓膜)，距离声源只有几米远，那么声球的曲率可以忽略不计，此时声音以平面波的形式传播。即使声源有一个特定的方向(就像大多数扬声器一样)，只要膜片厚度大于声音的波长，声音仍然会以球状形式传播。而波长短的声音在一定程度上保持其原始方向，在足够高的频率下它就可以形成声波束(我们将会在第6章谈到)。

声音的载体

声速只取决于介质的弹性和密度(见第1章)。在空气中，声速随湿度和温度的增加而增大[2]，但这仅仅是由于这些因素引起了空气密度的变化。表1给出了声音在一些介质中的传播速度。

介质	声速/米·秒$^{-1}$	与20℃相对湿度80%的空气中声速之比
0℃的氢气	1286	3.73
0℃相对湿度0%的空气	331.45	0.962
0℃相对湿度80%的空气	331.70	0.963
20℃相对湿度80%的空气	344.37	1
25℃的水	1493	4.33
25℃的海水	1533	4.45
钻石	12 000	34.85
花岗岩	5950	17.28
铁	5130	14.90
硫化胶管	55	0.16

表1　不同介质和条件下的声速

由于空气中的声速随着温度的升高而增加，而在白天，海拔较高处的空气比近地面处的空气温度高，所以高处声音传播的速度也更快一些。这种速度的增加使声音在较热的空气中向下弯曲(折射)，并在一定距离内返回地面，如图3所示。由于折射效应，有时声音在很远的地方比在近的地方听得更清楚。

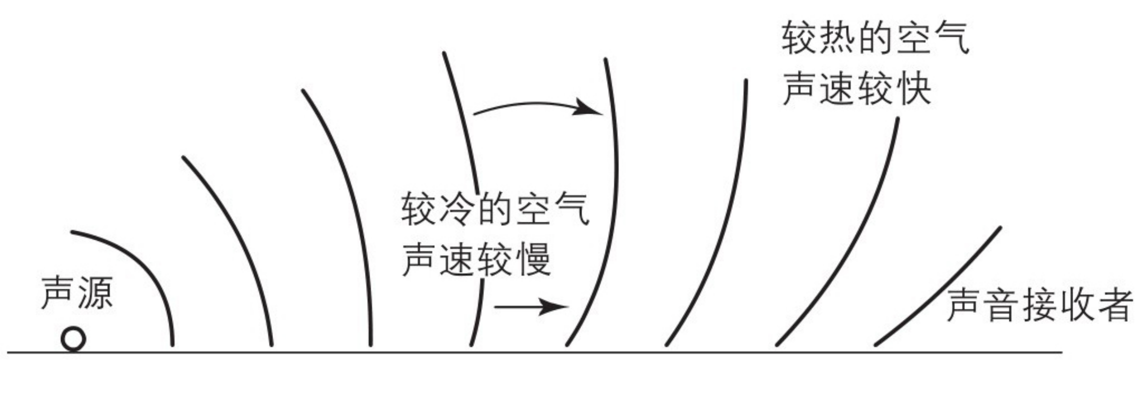


图3　当离地面较近的空气比较高的空气温度低时声音的传播

折射也解释了为什么在迎风情况下人们很难听见声音。在迎风情况下，风会使声波的传播速度稍微变慢一些，越往高处风速越快，声波速度降低得就更多一些，所以离地几米高处的声波的传播速度相对来说要更慢一些。声音从低速区域折射到高速区域，因此声波将远离地面和人的耳朵，朝上方传播(见图4)。

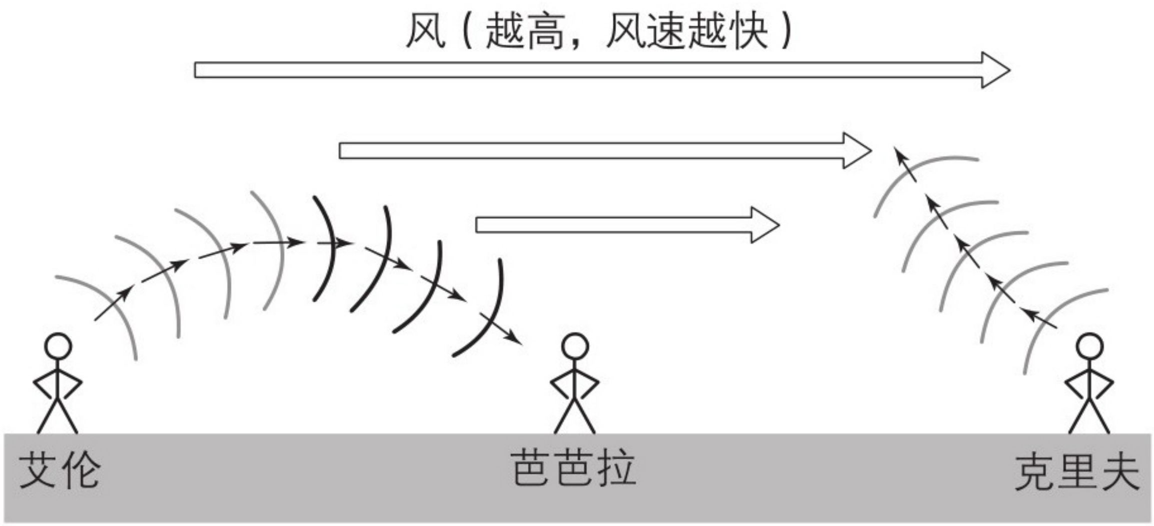


图4　芭芭拉可以听到艾伦的声音，但听不到克里夫的声音

无论怎么操作膜片，我们都无法使声音在膜片周围的空气中传播得更快。振动越快，产生的压力脉冲就越接近，它们到达某处(比如说鼓膜)时的频率就越高。也就是说，声音的频率会上升。如果一个人试图通过加快膜片向内和向外移动的距离来更用力地推动空气，那么脉冲中的压缩量和稀薄度就会增加，从而导致声压更高，听起来声音也就更大。如果我们强行让膜片的运动速度超过介质中的声速，那么在下一个脉冲形成时，前一个脉冲还没有来得及离开膜片，因此，它们就堆积成一个单一的、被称为冲击波的极高压脉冲，这也是音爆和响鞭声的成因。

快速地移动膜片并不是唯一可以用来增加声音频率的方法。如果扬声器(或其他声源)迅速接近你或你迅速接近它，压力脉冲就会以更高的频率到达你的耳朵，因为每一个脉冲都比它前面的脉冲在离你更近的地方开始向你传播，从而导致声音的频率上升。而当声源从你身边经过后，脉冲就会以更长的间隔到达你的耳朵，因为每一次脉冲的传播距离都比前一次要长，相应地，频率也会因此下降。这就是著名的多普勒效应。当你身边开过一辆超速的摩托车或当你听到跟在摩托车后面呼啸而过的警车的鸣笛时，就会注意到多普勒效应(见框2)。

框2　多普勒效应

对一个正在靠近观察者的声源，观察者听到的声音频率为$f_{obs}=(\frac{v+u}{v})f$。

其中，v是声速，u为声源靠近的速度，f是声音的频率。对一个正在远去的声源，我们只需要将公式中的“+”变成“-”即可。

与光类似，如果反射表面光滑且坚硬，声音就像光从镜面反射一样能够反射形成一个声源的像。所以如果你位于声源和反射表面正中间的某个地方，你从两侧听到的声音是差不多的(反射的一侧声音会稍小些)。当然，“光滑”是一个相对的概念，它意味着“表面凸起的大小比波长的长度小”。由于3千赫的声波波长比黄色光的波长要长100万倍，所以即便是粗糙如混凝土一般的表面也可以成为很好的声波反

射镜。凹面声波反射镜能将反射的声音聚焦。例如，在第一次世界大战期间，英国南部海岸建造了这种凹面混凝土声波反射镜，将飞机靠近时的声音聚焦到正在监听的士兵耳朵里。当声音在两个或两个以上的曲面反射器之间回响时，可能会产生一个耳语廊[3]，就像伦敦圣保罗大教堂的长廊一样。

声音能在任何两种介质之间产生界面反射，无论是在空气和混凝土、水和空气，还是地球上不同的岩层之间。反射声音的多少取决于两种介质声阻抗的差值，而声阻抗又取决于介质的密度和声速。声阻抗(见框3)与电阻相似，它测量的是声音在介质中传播的困难程度。它是声音诸多应用的关键。例如，软橡胶表面会吸收声音并将其转化为热量，因为软橡胶具有极高的声阻抗，潜艇上的隐形涂料就是利用了这一原理。但糟糕的是，橡胶的柔软程度与温度有关，所以，在20世纪80年代末，冷战时期的潜艇从北大西洋重新部署到海湾时，由于水温随地域变化而逐渐升高，使得潜艇无处遁形，从而引发了一系列相关研究的热潮。

框 3　声阻抗

介质的特性声阻抗计算公式为：$z_0=\rho_0 v_0$

其中，ρ_0为密度，v_0为声速。声阻抗的单位是瑞利（Rayl），下标“0”代表介质中的这些量都是“不受干扰”的，即介质中没有声音。

声音可以通过声透镜聚焦。声透镜通常由丙烯酸塑料制成，其工作原理是当波从一种介质传播到另一种介质时，只要它以一定角度入射介质之间的界面，就会发生折射。声波被折射的角度取决于它在两种介质中的速度之比，这就是斯涅尔定律(见框4)。

框 4　斯涅尔定律

$$\sin\theta_1/\sin\theta_2 = v_1/v_2$$

其中，θ 为声音入射方向与界面法线之间的夹角，v 为声速。

通常，声音比光更受关注的一个效应是声音能够在墙角和墙壁上弯曲，并在穿过一个开口后扩散开来，这种现象被称为衍射或散射(见图5)。

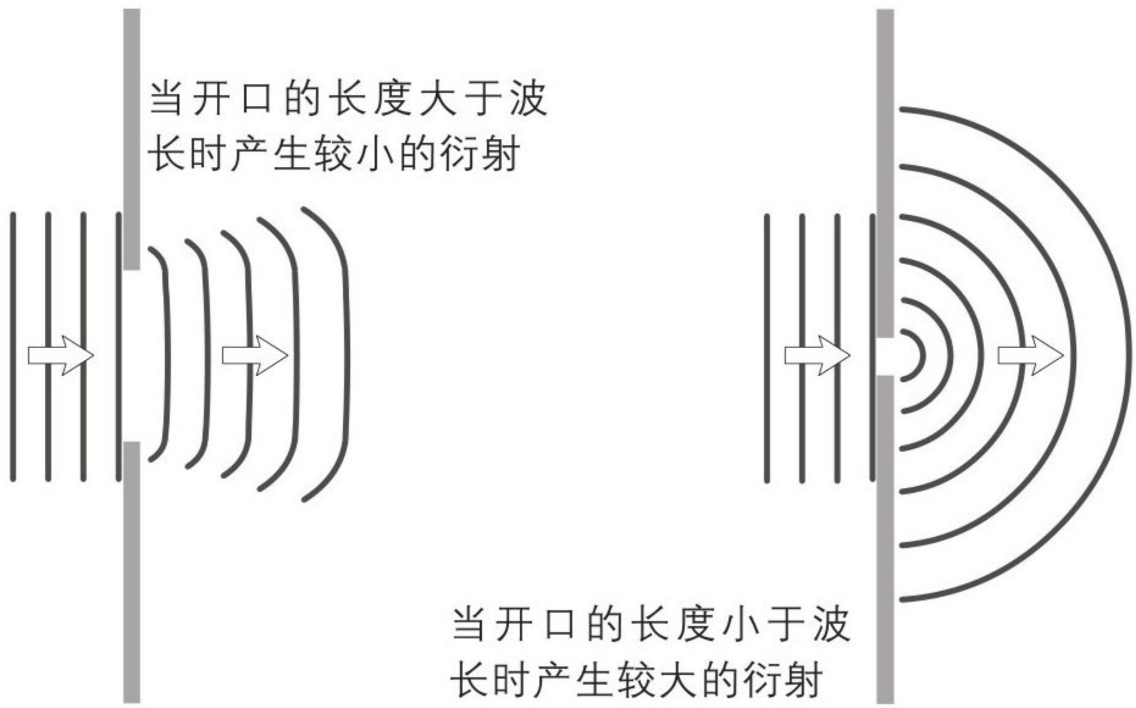


图5　衍射

波长越长，弯曲程度就越大。所以如果在声源(比如说一个乐队)和听者之间建一堵高墙，低音可以通过衍射效应绕过高墙返回地面，而高音则不能被听到(见图6)。这种消声效果对帮助我们评估户外常见声源的距离而言，是非常有用的线索。



图6　不同波长产生的衍射

当光线落在一系列间隔约为单个波长的平行线、条纹或凸起之上时，它就会发生衍射现象。由于波长较短的光将产生更大角度的衍射，这样的衍射光栅就会把白光分解成它的组成色，例如CD的背面就是通过这种方式将阳光衍射成彩虹色。由于纯音是一系列压力增加的有规则的“条纹”，它也可以充当衍射光栅，通过这个光栅散射的光波，长度约等于条纹之间的距离(该距离是声音波长的一半)。通常这里涉及的介质是晶体固体，如熔融石英。这种声光效应，即利用声波散射光，在水下和空气中都可被用作非摄动测量和成像工具(见图7)。

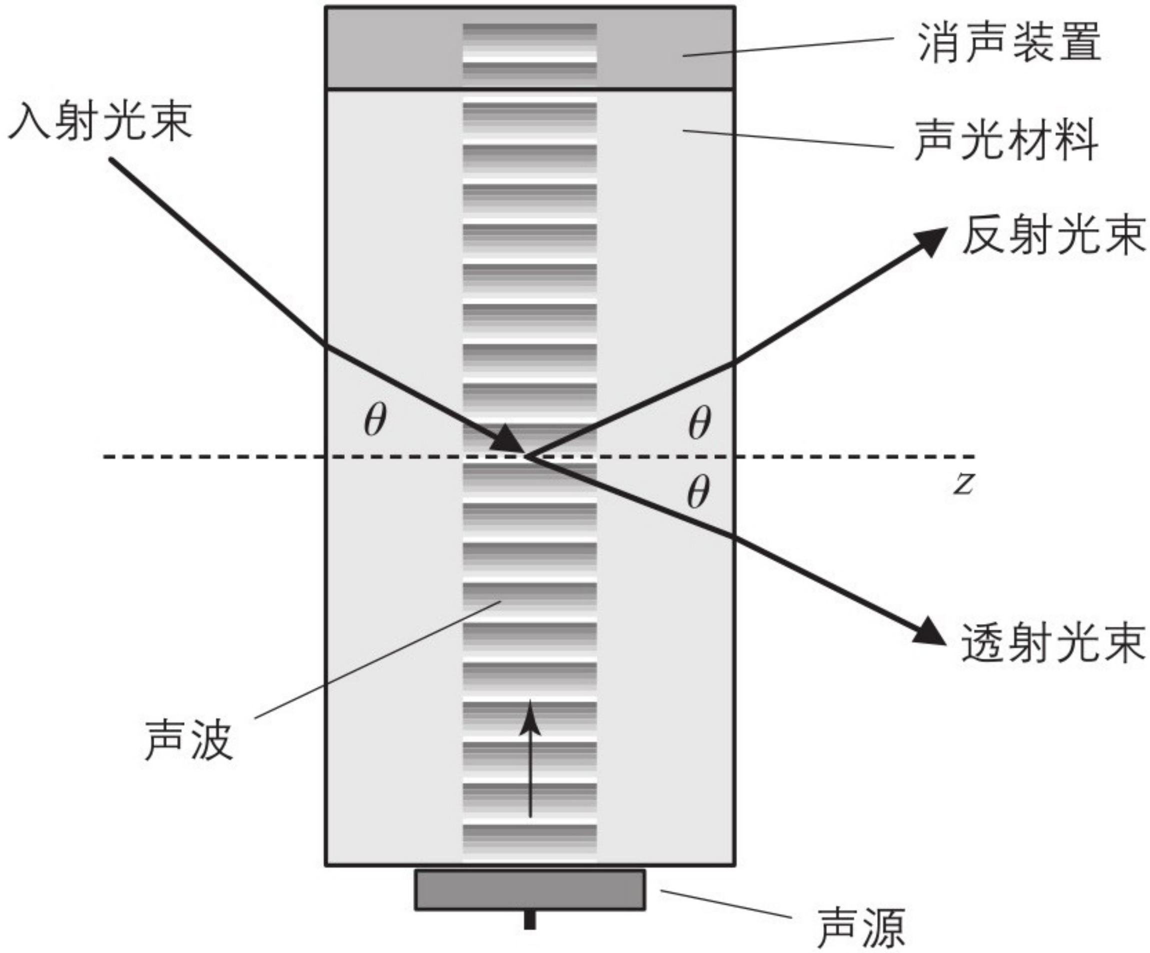


图7　声光效应

当来自多个声源的声音相遇、相容混合时，就形成了一个由嘈杂和安静区域组成的三维模式，称为干涉模式。安静的区域形成于一个声源的疏部与另一个声源的密部相遇的地方，这就是相消干涉；当疏部与疏部相遇，或者密部与密部相遇时，嘈杂的区域就出现了相长干涉（见图8）。

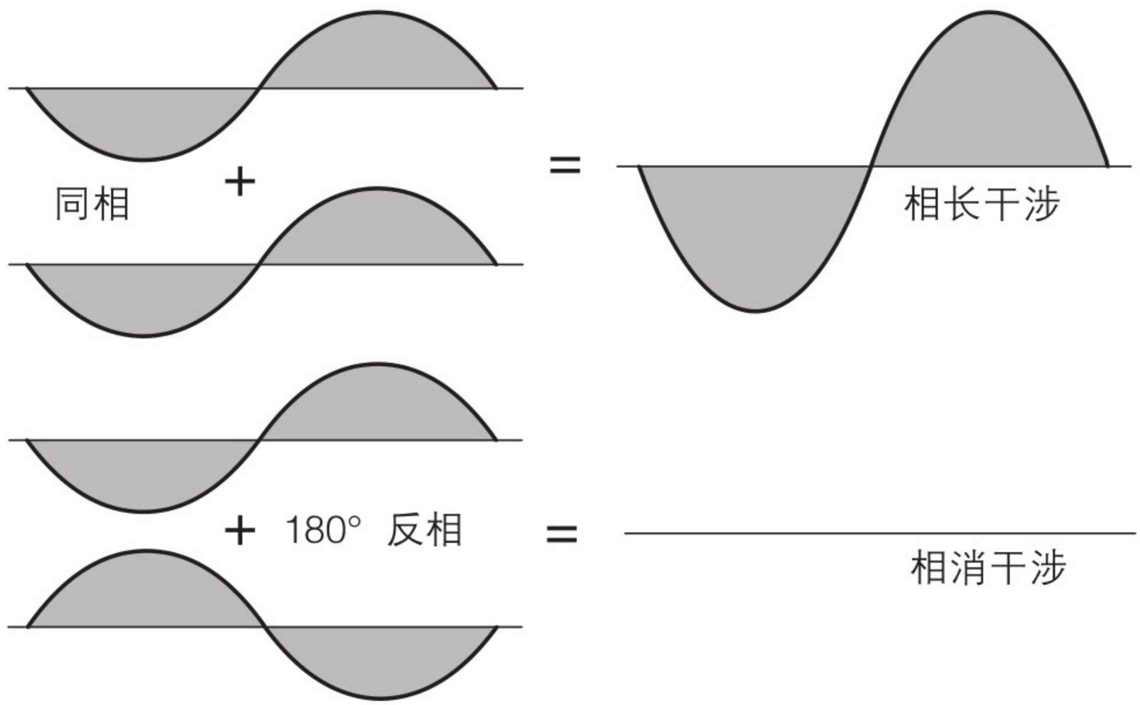


图8　相长干涉和相消干涉

干涉在立体声产生和噪声消除中发挥了重要的作用。它引入了声波的另一个表征参数——相位，即声波在空间和时间的某一特定点上的压力的高低。相位只在声波相互作用时才真正起作用。在上面的例子中，密部相互重合的一对声波(因此形成一个较嘈杂的区域)称为**同相**，而那些不重合的声波则处于**异相**。当波最大限度地偏离相位时，我们说它们处于**反相**。人类的听觉系统是无法识别相位的。

声音的能量

定义和测量声音的量的方法有好几种，每种方法适用的领域都不相同。如果测量的内容是听力或音乐，那么声压是最好的选择，因为它是与响度最直接相关的参数(尽管也不是很简单，但请继续读下去)。但是，在讨论声源的效率时，人们可能希望知道每秒有多少能量从声源中流出，即声音的能量。要描述特定声场对物体的影响，我们

感兴趣的参数是声强，即每秒打在该物体1平方米面积上的声音能量。音量是一种定义不太严格的度量方法，用于标记音频设备，但旨在模拟响度。

可听频率范围内的声波在通过空气时的吸收损失很少(每100米约0.25分贝/6音分，但随天气条件变化很大)。声音会随着距离增加而消失的主要原因是它们可以自由地向许多方向扩散，所以它们的能量会随着扩散越来越低，同时占据的体积越来越大。如果声源悬浮在自由空气中，其声音可以向各个方向传播(球面扩散)，则声压与接收者到声源的距离成反比。也就是说，如果从声源到测量点的距离加倍，声压就会减半。

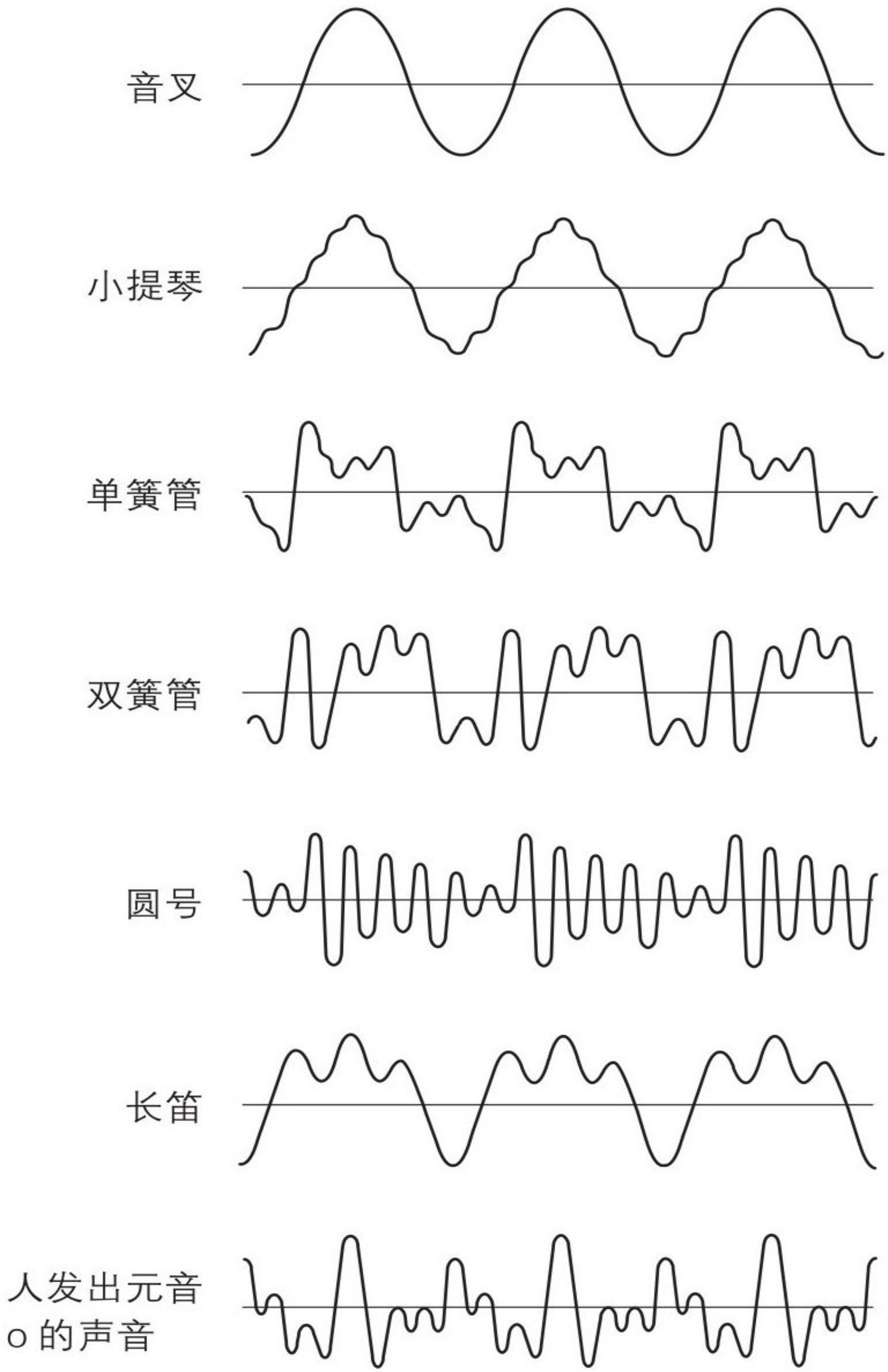
声音的强度下降得比这还快，它与距离的平方成反比。因此如果我们假设从声源到测量点的距离加倍，声音的强度就会下降1/4(1/22)；如果距离扩大为10倍，则强度下降到之前的1/100(1/102)。但如果声源在地面上，则声波呈半球面传播(见框5)，声压和强度的下降率变为上述情况的一半。换句话说，当距离加倍时，声音的强度大约下降到原来的一半。不过也有例外情况，当地面是一个完美的反射器时(一块大理石地板可以近似于完美的反射器)，强度的下降会比这更快，这是由于反射介质吸收导致能量损失。声功率只取决于声源，所以在任何距离上都是一样的。

框5　球面传播

$$I = P / 4\pi r^2$$

其中，I是强度，P是功率，r是测量点与声源之间的距离。

自然界中基本上不存在纯音，最接近纯音的可能是鸟类的歌声。真实声音的波形看起来各不相同，图9显示了基频相似的不同声音的压力变化。



音叉
小提琴
单簧管
双簧管
圆号
长笛
人发出元音
o 的声音

图9　不同声音的波形

困难的分贝

声音是最早被人类理解的一种能量形式。早在公元前300年，我们就已经知道它是可以穿过空气和水的某种形式的物理变化。但很久以后，声音最明显的特征——响度，才在一定程度上被真正量化。但是，直到2000多年后才出现的这种量化，并不能令人满意。

到目前为止，最广泛使用的量化声音的度量是分贝(dB，见框6)。如果两个信号在声压上相差1分贝，那么它们的声压比值约为1.2∶1，这恰巧是我们在理想条件下能听到的最小差异。10分贝的差异相当于声压比为3∶1，100分贝的差异相当于声压比为100 000∶1。

框6　分贝

以分贝计，能量为P_0和P_1的两个声音之间的强度之差为$10\log_{10}(P_1/P_0)$。

强度与压力的平方成正比，因此声压为p_1和p_2的两个声音之间的强度之差为$10\log_{10}(p_1^2/p_2^2)$，即$20\log_{10}(p_1/p_2)$。

分贝是1/10贝尔，贝尔这个词由传输理论中最常用的三个字母(β，ε 和l)结合而成，并且也是向亚历山大·格雷厄姆·贝尔(Alexander Graham Bell)[4]致敬。分贝并不是单位，它其实是比值，所以它可以描述一种东西比另一种强大多少。如果你愿意，你可以用它来比较一对加热器的输出，但这并不能告诉你它们实际上到底有多热。

要用分贝来描述一个设备的声音，重要的是要知道你在和什么进行比较。对于空气中的声音，我们将其与某种刚好能被听到的声音(相当于20微帕斯卡的压力)进行比较。当声音的大小以这样的参考水平给出时，就会加上“级”字，例如，**声压级**(Sound Pressure Level，SPL)。

所以，0分贝的声音是你能听到的最小声音的“1倍大”(也就是说，和你刚好能听到的声音一样大)，1分贝大约是12倍，2分贝是26倍，以此类推。所有的声学家都满意这个解决方案吗？不，他们并不满意。超声波工程师并不在意他们的超声波比“你恰好能听到的声音”大多少倍，反正本来也没有人能听到。他们喜欢的是功率，因此他们测量数据的单位是瓦特。与此同时，水下声学家理所应当地问道：“那么听力的阈值呢？当你的耳朵灌满了水，头上戴着橡胶耳机，这种阈值有什么意义呢？”又或者：“如果你是鲸鱼呢？这种阈值又有什么意义？”所以，他们的分贝建立在1微帕斯卡的参考压力上，因为这很好记，也很容易测量。于是现在我们有两种“分贝”，一种用于水中，另一种用于空气，这两种分贝对于相同的声音会给出不同的值。只要每个人都记得他们所使用的分贝的参考水平是多少就不是什么大问题，然而糟糕的是人们并不记得。

还有一个问题。很少有人关心一个物体究竟产生了多少声音，我们想知道的只是它的声音听起来有多大。声音听起来有多大取决于物

体离我们有多远，这似乎是显而易见的，但这意味着我们不能说汽车喇叭的声压级是90分贝，只能说它在一定距离上是这个值。通常情况下，即使是那些在教科书中非常流行的简易分贝图表也会出现这种错误，例如声称风钻的声压级为100分贝，而实际上是“如果在10米的距离(或其他距离)处测量，声压级为100分贝”。不难看出这种“偷懒”的表述到处都是，比如“安静的办公室”，我们知道它指的是你工作时安静的办公室，而不是走廊尽头或其他城镇的安静办公室。

还有第三个问题，即一个声源可能以任意一个、几个或多个频率产生声波。让我们暂且假设一下，声音的来源是一个扬声器，它的效率非常高，能把输入的所有电能都转换成声音。而且我们假设它有一个频率控制器，但没有音量旋钮。如果我们测量每秒从扬声器中流出的总声能(功率)，同时改变频率，那么功率当然会保持不变。同样地，扬声器的声压级在一定距离上会保持不变，这点用一个麦克风就可以证明(假设它在所有频率上都具有相同的灵敏度)。

然而，这和你的耳朵告诉你的完全不同。如果扬声器在20赫兹时刚好可以听到，它的音量会随着频率的增加而提高，直到大约4千赫时，它的音量会(非常粗略地)提高200倍。在更高的频率下，它会再次变得安静，最终在8～20千赫之间消失不见。你到底听不到什么频率的声音，取决于你的年龄，以及你在过去几十年里有没有好好保护你的耳朵。

在实践中，声学家对麦克风所组成的电路的响应进行加权，从而使系统表现得像耳朵一样，也就是对4千赫左右的频率最敏感。频率加权麦克风是**声级计**(Sound Level Meter， SLM)的核心。实际应用中有很多不同的权重可供选择，甚至可以选择适合狗的权重。目前最受欢迎的是A权重，它近似人耳在中等音量下的反应。因此，对人类重要

的分贝通常是A权重的，记为dBA，其全称是“A级加权声压级(以分贝记)”。

声级计会受声音时长的影响。这很重要，如果一个声音持续的时间短于0.1秒，它听起来就会更安静，因为人的听觉系统会将声音持续时间之内的能量积累起来去感受。

更复杂的是，音量的大小也取决于其来源的性质。例如，人们非常不喜欢飞机的声音，一般来说，他们认为飞机的声音就和实际上要大5分贝的无特征声音一样令人讨厌。相反，人们更喜欢火车的噪声，他们甚至觉得火车声音的烦人程度和实际上低5分贝的无特征声音一样。这些反应如此明确，以至于许多涉及飞机或铁路噪声的规划应用程序将其数字调整了5分贝，这种调整被称为“飞机惩罚”和“铁路奖励”。这意味着，没有任何仪表可以真正测量建筑师、房主、噪声活动家、嘈杂的机器购买者和声学家真正需要知道的东西——音量有多大。

考虑到所有这一切，用高精度声级计测量声压级几乎没有意义。大多数声级计测量10千赫声音的精度为± 1.4分贝(称为2类仪表)。即使在实验室工作中，测量精度为± 1.1分贝基本上是足够的(由1类声级计提供)。比准确度更重要的是遵守标准测量程序，包括通过与标准测量麦克风的比较来频繁地校准声级计。

尽管响度具有复杂性，而且会根据声源和用户的不同而产生变化，但科学家通过精心选择声音，然后对大量被测者的反应进行调查，大致确定了响度与声压级之间的关系，并在此基础上定义了单位，也就是方(Phon)。方被定义为与1千赫单音的声压级具有相同的值，因此声压级为10分贝的1千赫音调的响度级别为10方。但是一个和10方相同响度的50赫兹声音，其声压级是73分贝，因为我们的耳朵对

50赫兹的敏感程度远低于1千赫的声音，一个50赫兹的声音需要比一个1千赫的声音高63分贝才能听起来同样响。

响度只是众多心理声学测量的指标之一，也被称为声音质量参数(此处“质量”指的是“性质”而不是“好坏”)。响度是目前最常用和发展最好的指标，其他的指标还包括锐度(单位是acums)、粗糙度(单位是aspers)、起伏度(单位是vacils)和柴油度(diesel)(柴油度没有单位，不同的汽车只是根据人们认为它们的声音有多“柴油”而进行主观排序的)。从最后一个名字能看出，这些指标主要是由汽车行业开发的，其目的是使门的咣当声、发动机的声音，甚至指示器的声音听起来更有力、男性化、可靠等。原则上讲，如果电子产品和噪声源都能用这些参数来描述的话，对用户是非常有用的。

声音质量这个话题是心理声学学科的一部分，即研究声音的心理效应，它本身可以被认为是现在所说的声音研究的一个组成部分。声音研究涉及各种声音在历史和不同文化中是如何产生和消耗的。关于这类主题的研究工作自20世纪40年代以来一直在进行，而自20世纪90年代初开始，这类研究的数量也在大大增加。

驻波

在声学史上，实现声音的可视是人们不断尝试的目标。在18世纪80年代，恩斯特·克拉德尼(Ernest Chladni)研究了金属板被小提琴的琴弓划过发声时振动的方式。撒在板上的细粉会偏离振动强烈的区域，并聚集在静止的区域。因强烈振动而没有粉末的区域对应于波腹(如图1中的波峰或波谷)，而静止且有粉末的区域是节点，即没有压力变化的点(图1中的曲线与轴线相交的地方)。

克拉德尼之所以能够以这种方式“看到”声波，仅仅是因为声波没有在空间中前进。它们是静止的，或者是“停驻的”声波，即驻波。对于驻波，图1只表示波的压力随位置的变化，而不是特定点的压力随时间的变化方式(驻波中任何一点基于时间的压力示意图都是一条水平线)。

这种原理在下面的例子中会展示地更清楚。假设有一根长12厘米的管子，一端开口，一端封闭。如果有人从开口端吹气，管内就会形成驻波。在这类波中，由于空气与闭合端管壁有摩擦，那里的空气无法移动，所以这一点是波节点。最简单的一种驻波是这样的：空气分子离闭合端越远，运动幅度越大，在开口端幅度达到最大(这一点称为波腹)。在这种驻波中，波长的1/4与管长相等，所以它的波长为4× 12 = 48厘米。如果你吹得足够用力，管中就会形成一系列其他形式的驻波，每个驻波都有节点和波腹分别出现在管的两端，如图10所示。这些其他波的波长是第一种驻波的简单倍数，这样的一组波被称为谐波。

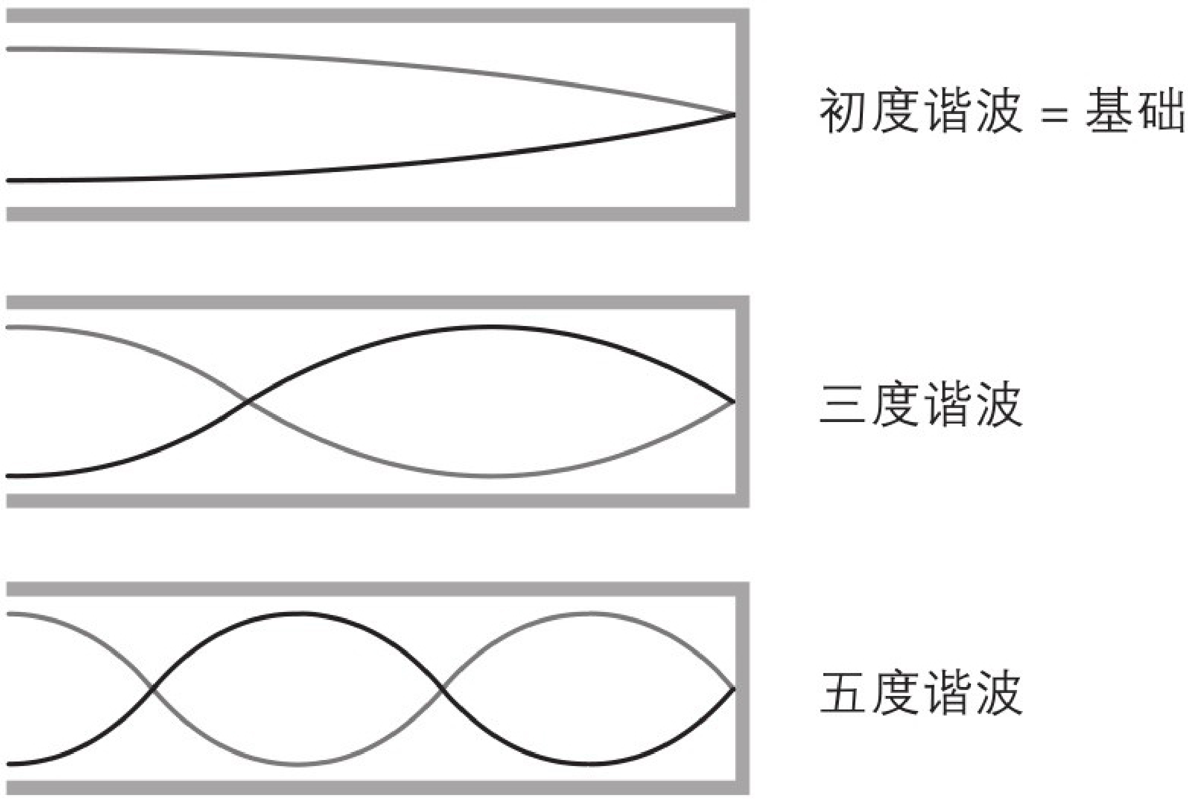


图10　一端开口管中的驻波

与上述的管子类似，在任意充满液体的腔或任何坚硬的物体里，都存在非常容易被激发的特定波长的声音。这被称为共振模式(或简称为共振)，共振的主要模式是可以预测的，因为它们只依赖于尺寸。例如，如果把一根12厘米长的杆子的两端固定，猛击它将产生24厘米波长的声波，以及波长为12厘米、8厘米、6厘米、4厘米的声波，还有所有其他节点间距能达到12厘米[5]的声波，这同样是一组谐波。

一个12厘米长的盒子里的空气或水也会产生上述所有波——在这种情况下，流体的“末端”靠近盒壁，在那里摩擦力阻止了流体的自由运动。盒子还会产生与它的高度、宽度和对角线相对应的波族。

共振是室内声学领域的一个主要研究问题，也是大多数乐器的基础。如果乐器的一端是开放的(比如风琴管)，开口的一端就是腹点，其基频的波长会是相同长度的封闭管波长的两倍(实际上波腹在管道的末端之外形成，需要进行末端校正，请参见框7)。

框 7　末端校正

闭合管$\frac{\lambda}{4}=l+c$，开口管$\frac{\lambda}{2}=l+2c$

末端校正值c约为管半径的 0.6 倍。

通常，最低共振频率是最强的。然而，如果给一个乐器提供大量的能量，它可能会产生一个八度共振，甚至更高到两个八度。例如，如果吹得足够用力(“超吹”)，长笛就可以做到这一点。

共振在我们的生活中随处可见。当你敲击盘子、杯子或叉子的时候它们会发出声响，只要你别把它们握得太紧，否则声音会被抑制住(音叉在紧紧握住的时候仍然可以产生共鸣，是因为音叉有两个相同的尖头叉子且移动方向相反，这样共鸣就在把手处相互抵消，因而没有在把手那里产生振动)。利用共振，我们可以检查陶器是否有裂纹。如果没有裂纹，陶器材料中每一毫米的运动都会紧随相邻毫米的移动，让波通过，就像墨西哥人浪[6]一样，这表明陶器确实是完好无损的。但是，即使是一个非常细小的裂纹也会将相邻区域分开，此处的拖曳和摩擦也会抑制共振，产生反常的“叮当声”。

如果一个力以与该物体的共振频率相同的频率作用在物体上，耦合的效应就会非常明显，因此吉他弦会呼应房间另一端的拨弦声发出

声音，或者电视机的一些部件伴随着节目声音突然发出恼人的嗡嗡声。

在声学的几个领域中，一个重要的效应是**亥姆霍兹谐振**(Helmholtz resonance)，任何听过吹瓶口时产生的音调的人都很熟悉这种效应。只要是有开口的空心物体或腔体都可以作为亥姆霍兹谐振器(见框8)。如果一股气流从开口上吹过，一些空气会进入腔体，增大腔内的压力。压力又把空气推出去，就像钟摆一样，让这个空气“过度冲出去”，留下一个轻微的低压环境，这样就会吸入更多的空气，循环往复。这种有规律的循环包含了共振频率的声波。如果把这个频率的声波供给谐振器，它就会发出非常强的声音。

框8　亥姆霍兹谐振器频率

$$f \approx \frac{v}{2\pi}\sqrt{\frac{S}{lV}}$$

其中，v 为声速，S 为瓶颈面积，l为瓶颈长度（包括末端校正），V 为腔体体积。

绘制声音

驻波是声波的一个小子集，大多数情况下，波中的高压区和低压区在空间中运动(这种波被称为前进波或行波)。如果你想“看到”行波，就必须记录气压随时间的变化。亚历山大·格雷厄姆·贝尔是第一个尝试这种方法的人。1874年，他设法从一具尸体上弄来一只耳朵，给它涂上油以保持其柔软，并在鼓膜上绑上一根细稻草。稻草的另一端可以在一块带着煤灰的玻璃上画一条线，这样一来，当有人对

着这个耳朵大喊大叫时，玻璃就会随之移动。这条摇摆不定的线是声波的第一次记录，这个装置被称为人耳记音器。为了减轻那些需要制作这种装置的人的负担，后来的版本不再使用死人的耳朵，取而代之的是金属振动膜。

然而，记音器对于声波的实际测量并没有什么用处。实际的测量最终由在20世纪30年代开发的**阴极射线示波器**(cathode ray oscilloscope， CRO)实现。CRO可以设置不同的时基，这样高频的声音就可以在屏幕上传播，而低频的声音则被压缩，于是就可以看到它们的波形，由此可以读出它们的波长并确定它们的频率。

今天，CRO在计算机上被广泛使用。然而，二维图仍然只能显示声音的一些特征。大多数声波的频谱和压力变化都很快，只有在一种叫作声谱图的三维示意图上才能正确地同时显示出来。这种声谱图需要依靠计算机才能生成。在声谱图中，屏幕上方的高度通常代表频率，亮度或颜色代表声压(或强度)。在其他情况下，可以在屏幕上显示三维形状，结果通常类似山脉(见图11)。



图11　声谱图

解码声音

能够看到声音，意味着人们可以对很多与它相关的信息进行定性，也可以对显示结果进行粗略的测量，但是很多时候我们还需要声音的精确定量信息(例如为了消除噪声或改进乐器的设计)。为此，我们需要进行数学分析，这是最广泛和最基本的分析方法之一，是基于约瑟夫·傅里叶(Joseph Fourier)在19世纪所做的工作而产生的。

傅里叶意识到任何周期函数(以稳定速率重复的函数)都可以通过将一系列正弦波(现在称为傅里叶级数)加在一起来构造，他还想出了一种方法来确定这个级数的元素(项)是什么。从数学上讲，傅里叶级数是由一系列正弦和余弦组成的——考虑到余弦只是一个从最大值而不是从0开始的正弦波，所以我在这里只提到了正弦波。如图12所示，三个正弦波就可以近似一个方波。为了使后者的侧边更垂直，必须添加更高频率的音调。方波听起来像咔嚓声，傅里叶分析表明，突然(声级快速增加)的咔嚓声会包含一些非常高频的成分。

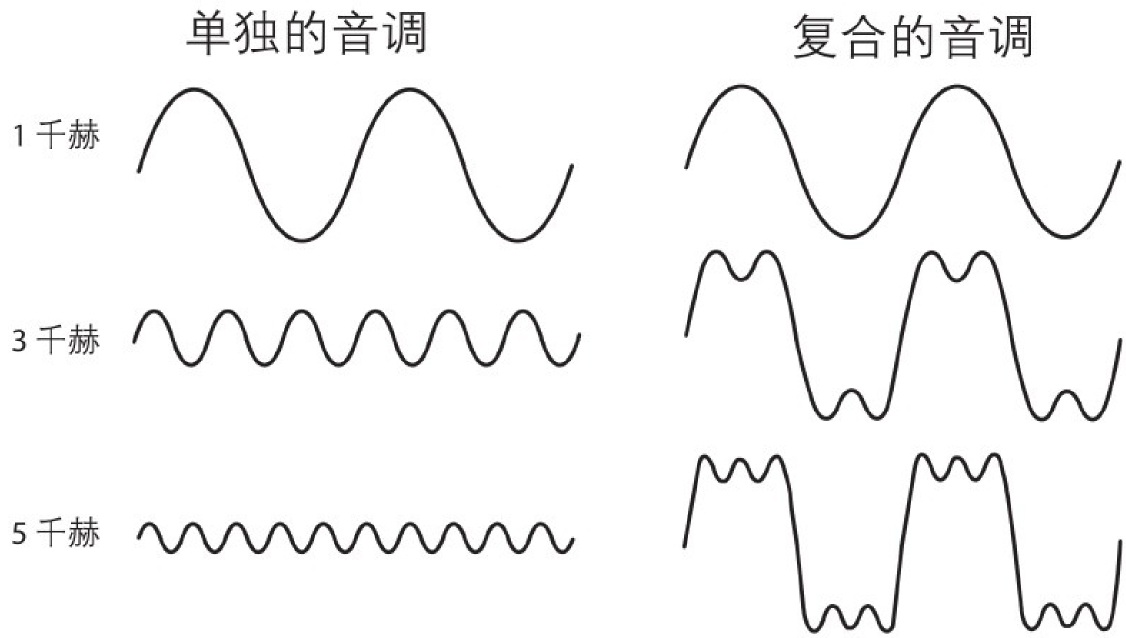


图12　用正弦波求和来近似一个方波

傅里叶的原始工作只适用于周期波，但它进一步发展为一种被称为傅里叶变换的方法，可以用来处理非周期波。**快速傅里叶变换**(fast Fourier transform， FFT)是一种计算信号正弦波分量的高效数学方法。当加入这样的波时，人们就必须考虑它们的相位。在一个波长内，声波的声压从零(与周围空气压力相等)上升到最大值，然后下降到最小值，然后又上升到零。这类似旋转的轮子边缘某一点的垂直运动轨迹，因而相位可以用角度的方式来描述，即从0° 开始上升到最高相位的90° ，再落回水平位置180° ，接着降至最低位置270° ，最后回到360° (与0° 是一样的)。

所有真实的声音都会随着时间的推移而改变，所以转换成正弦波的过程必须频繁地重复。这种对声音的时变频率分析有许多应用。例如，组成一个人声音的声波的某些参数对那个人来说是独一无二的，因此，这些参数可以作为“声学指纹”(即声纹)，并可以由机器进行自动语音识别。

相反，由于每个单词都有独特的发音(除了同音异义词，比如“sew”和“so”)，无论由谁来说这个单词，理论上机器都可以自动识别出来。虽然不同的人有发声差异，但某些特定的因素只是略有不同，或者是可以预测的。这也是为什么(在一定条件下)不管说话者是谁，我们都能够识别一个单词的含意。

然而，自动语音识别离完美还有很长一段路要走，主要的问题在于判断一个单词的结尾和下一个单词的开始位置。这个问题很难，试着听自己说“面包和黄油”，你听到的可能是像“brembudder”这样的词[7]，没有任何的停顿。人类之所以能够如此容易地识别单词，是因为我们听到的声音模式只是所说内容的表征之一，正如第4章将要解释的那样。

合成声音

由于任何声音都可以被拆分成正弦波，因此任何声音都可以从正弦波合成而来。从声音中生成语音的合成器已经问世多年，而且效果比识别器要好得多。然而，在实践中，通过将预先录制或预先生成的声音片段组合到一起来生成语音通常要容易得多，这是一种被称为语音编码的技术。

目前的电子系统几乎可以合成任何声音，不管它是否原本就存在于自然界中。比如怪异的谢泼德音调(Shepard tone)，它是由逐渐降调又渐出的音调组合而成，其间会有更高的音调渐入，然后也开始下降，给人的印象是声音不断下降却又不会下降。然而，通常人们并不需要新的声音，而是需要现有声音的改进版本，例如一场没有噪声的音乐表演。在20世纪60年代，对流行音乐制作人来说最著名的电子产品就是电子琴，其加载的预录元素是流行音乐常用的非语言声音。这

种乐器上加载的库中有录制在磁带上的小段声音，演奏者可以迅速用他们想要的频率和音量来播放任何片段。

选择声音：滤波器

最常见又最简单的修改声音的方法是过滤，即使用电子电路或软件删除或减少选定的频率范围。高通滤波器去除低频，低通滤波器处理高频，带通滤波器则同时去除低频和高频。曾经有一种很常见的可变滤波器是图形均衡器，它是高保真音响放大器上的一系列滑块(约七个)，可以抑制选定的预置频率范围。简单的“音调”控制同样可以使高(高音)或低(低音)频率变得安静。

各式各样的其他功能也出现在电脑声音艺术家或工程师的工具箱中。例如，有些软件可以添加混响或回声，创造一个人造声景，或在用扬声器播放之前，用这种实时变化的频率改变之前预录的流行歌曲的声音。这是卡拉OK系统的基础，在卡拉OK系统中，歌曲的音调可以被升调或降调，以匹配用户认为最容易唱的方式。

[1] 牛顿摆(Newton's Cradle)是一种由5个紧贴的同样质量和大小的金属小球构成的玩具，当拉起一端的小球使其摆动撞击其他4个时，另一端的小球将会继续摆动，而中间的3个小球保持不动。

[2] 原文为decrease with increasing temperature，疑有误。

[3] 当两个人贴墙站立时，即使一个人只是对着回廊壁窃窃私语，走廊另一端与他遥远相隔的人也可以清楚地听到。

[4] 这里Bel与物理学家贝尔(Bell)的名字谐音且拼写方式相近，故有致敬的含义。

[5] 原文为24厘米，疑有误。例如波长16厘米的声波，不能由文中的长杆生成，但其节点间距可达24厘米。

[6] 球迷在看台上有序地举手站起再坐下。

[7]　这里由于连读现象使得原来三个词的短语“Bread and butter”听起来像一个词。在汉语中类似的例子有“这样子”，由于连读听起来像“酱紫”。

03 和谐之声 Sounds in Harmony

音符是什么

“音调”(tone)和“音符”(note)[1]这两个词分别反映了声音的主客观性质，音调是具有特定频率的声波，音符是其给人留下的主观印象，对应特定音高。

除了音高，一个音符还具有时值、响度和音色等特征。音色是迄今为止识别乐器最重要的特征，也是情感内容的主要载体。实际上，由于音色极为复杂，它与其他特征相比有明显区别。这有点像根据性别、身高、体重和指纹来定义一个人，前三个参数都可以明确到某项具体值，但是对一个人指纹的完整描述是非常复杂和多维的。同样，只有音色是特定的乐器或人所特有的特征，所以它是大脑用来识别声音唯一性的东西。

指纹侦探没有这种担忧，但对大脑来说音色是个问题。音色是一种动态特性，也就是说，它会随着时间而改变。铙钹、钢琴或鼓的音色从开始到结束都有很大的变化(这种变化称为流)。特定乐器或扬声器的音质也会随着音调和响度而改变。例如，大提琴的高音与低音相比有着不同的音质，同样的一个字吼出来和讲出来相比并不仅仅是声音大了一些，而一个男高音歌唱家在唱出他能达到的最高音时比唱低音听起来更单薄。幸运的是，我们可以依靠高度进化的听力，专注于声源发出的第一个声音，从而理解这一切。

史前议程

我们对音乐的欣赏和我们对噪声的厌恶可能源于进化的压力，这种压力塑造了我们远祖的听觉系统。他们听到声音之后的首要任务是识别它，且没有时间去纠结。弓箭弦的响声、雪崩的隆隆声、正在接近的滚滚蹄声或蛇发出的咝咝声，只有在听者反应迅速的情况下才是有利的声音，因此听觉系统将其分析能力集中在听到声音的最初时刻。这就产生了一个令人惊讶的结果，即一件乐器在最初几秒内发出“突发”声，尽管通常是不和谐的，听起来也不像乐器的“稳态”声，但正是“突发”声使我们能够识别出那件乐器是什么。一段经过编辑，删除了所有突发声音的录音音乐听起来很奇怪，我们很难识别出发声乐器。多年来，这一事实阻碍了人们制做更符合听觉规律的乐器。

通常情况下，乐器，甚至乐曲都能在一秒内被辨认出来。曾经有一个相当受欢迎的电台(后来又上了电视)智力竞赛节目叫《 听歌识曲》(Name That Tune)。你很容易就能识别一段你熟悉的录音，即使只是听到开始的一点声音，你也足以识别它。同样，一些声音会立即告诉听众整首曲子的味道：一个和弦可能有一种“墨西哥”的味道；几下小提琴弦的拨弦可能听起来有“乡村风”；一个尖的风笛音有“苏格兰”的风味；或者几个音符就描绘出“凶猛鲨鱼在逼近”。

我们对这些简短提示的反应主要是情绪上的，这与它们最初作为危险警告的作用相吻合。狮子的吼声或安全搭扣松开时发出的咔嗒声会立即引起本能反应，使人释放肾上腺素，为逃跑或战斗做准备。事实上，我们对身边突然出现的巨大声响的瞬间反应是远离它，如扬声器发出的尖叫声会让我们立刻跳开。即使是连续的声音，如果足够大，也会令人反感。因此，基本上没人愿意从风钻旁边经过。

与我们对突发声音的关注相反，我们会渐渐失去对持续声音的关注，这导致一个看似矛盾的现象，即恒定音量的声音在持续几十秒后

听起来变得安静了。在自然界中，这些声音听起来可能像微风一样无害。但是，如果这种长时间存在的声音突然停止，这种戛然而止就像声音刚出现时一样引人注目。

从声音研究的角度来看，听音乐的行为受到了相当大的关注。例如，音乐学家和哲学家彼得·森迪(Peter Szendy)认为，听音乐的重点是与其他作品、表演者、乐器以及其他听众的比较。他的观点是，由于听音乐的行为不断地以这种方式“占用”了其他东西，所以我们永远无法完全把握一段音乐的本质。

学习如何唱歌

对许多人来说，音乐之声远比其他类型的声音重要得多，以至于许多人几乎无法抑制自己想要突然放声歌唱的冲动。你通过唱《飞越彩虹》(Somewhere Over the Rainbow)来愉悦自己，可能在一定程度上也愉悦了周围的人。朱迪·加兰(Judy Garland)当然唱得更好，这一点毋庸置疑。但她是如何做到这一点的？虽然你的听众可能不喜欢这首曲子，但他们肯定能听出来，因为即使很多音符是错的，它们至少会按正确的顺序在音高上上下移动，并用他们熟悉的规律改变着音长。

音符和休止符(即音符之间的间隙)的长短由它们的形状表示，定义为“完全的音符”(全音符)长度的几分之几，如表2所示。

记号	名称	相对时长	休止符
𝅝	全音符	一个全音符	𝄻
𝅗𝅥	二分音符	半个全音符	𝄼
♩	四分音符	四分之一个全音符	𝄽
♪	八分音符	八分之一个全音符	𝄾
𝅘𝅥𝅯	十六分音符	十六分之一个全音符	𝄿

表2　音符和休止符

通常，音符没有标准的持续时间。尽管在某些情况下，作曲家在其乐谱上会标类似“♩=66”的记号，这意味着“一个四分音符持续1/66分钟”。然而，通常情况下，一首乐曲速度（专业术语为tempo[2]）的唯一标记是一个有丰富含意的短语（通常是意大利语），比如行板（以步行的速度）或从容的快板（快但不太快）。也许是因为节奏的概念在生活中不适用，我们对它的判断力很差——无论是业余爱好者还是专业人士，能发现乐曲里大约4%的节奏差异，就是极限了。

除非你是一个训练有素的歌手，或者天生的好歌手，否则你可能会唱出一些稍微“平一些”（也就是说，比本应该唱的音要低一点）或“尖一些”（比本应该唱的音要高一点）的音符。但这里的“应该”是什么意思呢？这并不意味着“唱对了音”。你在唱“在”字的时候是300赫兹还是333赫兹并不重要，重要的是你要得到一个正确的频率比。所以，如果你用300赫兹唱“在”，就必须用600赫兹唱“哪里”，也就是高一个八度，要不然人们听了就会直皱眉头。

因此，如果你在用钢琴给自己伴奏，唱的调却和钢琴不同，那听起来肯定是不合调的。对钢琴家来说，“在”是一个“Fa”。一架88

键的钢琴上有8个“Fa”键，任何一个“Fa”键都可以用在这里——如果选择第四个“Fa”，钢琴就会发出349赫兹的音符。

如果你真的发现自己能“自动”与钢琴合调，那么你可能拥有一种罕见的特质，即**绝对音准**(absolute pitch，也叫绝对音高)。拥有绝对音准的你只需要唱出与钢琴相同的音符即可。但如果你回到两个世纪以前，那时钢琴的音调和现在不一样，那你也就无法和它们保持一致。如果你想和当时绝对音准的歌手合唱一首无伴奏的二重唱，其结果听起来可能并不令人满意。

不过，从某种意义上说，你的音准比你19世纪朋友的要好。如果你穿越两个世纪去未来，你的绝对音准就可以和其他人相比较，但19世纪的人就不一样了。这是因为，1939年在伦敦举行的国际标准组织(International Standards Organization)会议上，许多国家都同意将特定音符规定为特定频率。同许多其他国际协定一样，产生这项协定是因为长久以来的走调情况已经变得让人无法容忍。

由于乐器制造者、音乐家和作曲家在不同的国家基本上是独立发展起来的，所以他们都采用自己的音高标准。在拥有多个音乐中心的国家，音高甚至因城市而异。1780年，一个拥有绝对音准的柏林人在维也纳也会走调。在一个城市制造的小号在另一个城市就会走调，而且小号一旦做出来就不能再调音了。因此，随着音乐家的国内和国际旅行变得越来越普遍，成功的乐器制造商想要将他们的客户群扩大到更远的地方，对国际标准的需求变得更加紧迫——因此就有了这次会议，会议举办的时间甚至定在战争前夕。

绝对音准并不是与生俱来的天赋，必须靠学习掌握，而且得在6岁之前学习，通常是通过记住乐器上的所有音符。从表面上看，这几乎是一项令人难以置信的记忆壮举，而这种记忆技巧对成年人来说已经是不可能习得了。有些孩子似乎更容易学习绝对音准。在中国和越南

拥有绝对音准的人要高于欧洲或美国，这可能是由于东方语言中的词语对音调有更高的依赖性。

对我们这些缺乏绝对音准的人来说，我们记住的是音符之间的差异(音程)，而不是音符本身。因此，无论是低音提琴还是小提琴，我们都能立即识别出来其演奏的熟悉曲调。

为什么我们喜欢我们所做的事情

到目前为止，音乐中最重要的音程是八度，即一个音符的频率是另一个音符的两倍。八度的重要性有一个事实支持，在我们今天大多数人使用的A到G的记谱系统中，相隔一个八度音程的音符由相同的字母表示。为了区分每个音符的不同，一般会要求使用数字下标(所以钢琴上的第四个F是F_4)。

一个八度的音程是和谐的，我们听到和谐音程，即和音的感觉就是听觉上的愉悦。除两个相同的音符外，八度是所有音程中最和谐的音程。

虽然我们知道大脑对谐音的反应会受神经刺激，但对和谐音程听起来很协调的原因还不是很清楚。也就是说，对这类声音的每个组成部分做出反应的神经元倾向于彼此同步它们的放电速率，因此，和谐的音程对我们来说显然很重要。为什么我们觉得和谐音程令人愉快，而不和谐的音程就没有这种感觉呢？事实上，尽管我们通常不喜欢单独的不和谐的声音，但从古希腊时代起，作曲家和他们的观众却渐渐喜欢上了不和谐的声音。西方音乐史上一直有人不断采用更多的不和谐的声音，这就是为什么听斯特拉文斯基(Stravinsky)的《春之祭》(Rite of Spring)或莫扎特(Mozart)的《弦乐四重奏》(Dissonance Quartet)不像过去的那样令人害怕。然而，如果按照人们的感受对音程

进行排序，其结果常常与地理或历史因素无关，所以和谐是客观存在的。事实上，有两件事值得注意：

1. 我们的耳朵还不够“粗糙”。当耳朵接收到两个频率相近的音调时(大约5%的不同，或两三个半音)，我们会认为听到单个不太精确的音符。

2. 与谐波系列的相似音调，也就是说一组具有1千赫、2千赫、3千赫、4千赫等频率的音调也很和谐。

更多的音符

音乐创作不能仅仅依靠八度音阶。首先，人类的听力范围不超过10个八度。其次，演奏出这样一个音域也是一种挑战，小提琴和吉他能演奏大约四个八度，天赋异禀的人能演唱五六个八度，钢琴能演奏七八个八度。此外，如果每首曲子都是最纯粹的和音，很快就会变得单调乏味。

八度音阶被定义之后，毕达哥拉斯和其他探索者就开始在八度音阶上添加音符，开始了一项长达千年的研究项目。尽管毕达哥拉斯研究的是单弦，但最容易的方法是用几个长度可变的弦，并要求它们都具有相同的张力和粗细，且由相同的材料制成。

在八度音阶之外，第二和谐的一对音符是由一对长度比为2/3的弦组成的。其他三个简单的弦长比产生了其他听起来相当不错的音符组合，如表3所示。

音程名	短弦与长弦的长度比	与第一个音符的频率之比	和谐度的顺序
一度	1	1	1
二度	8/9	$1\frac{1}{8}$	6
三度	4/5	$1\frac{1}{4}$	5
五度	2/3	$1\frac{1}{2}$	3
四度	3/5	$1\frac{1}{3}$	4
八度	1/2	2	2

表3　和谐音程

如果我们按照表3中定义的弦长制作一个竖琴，并要求弦的粗细、张力和材料都一样，我们会很容易得到简单、听起来令人愉悦的和弦。事实上，任何一对或大量这样的弦在一起弹奏，听起来都不会太糟。

这个系统被称为五声音阶。它是由世界各地的许多古代音乐家独立发现的。至今仍很流行的《奇异恩典》(Amazing Grace)、《我的女孩》(My Girl)和《我射杀了警长》(I Shot the Sheriff)等乐曲都是按五声音阶创作的。钢琴上的黑色琴键是一系列五声音阶，所以如果你随意按下它们，听起来不会像随便乱按白色琴键时那么糟糕。

更多更多的音符

人们虽然喜欢和谐的音乐，但同时也喜欢带一点点不和谐的音乐，五声音阶对那些想要挑战听众、突破界限的作曲家来说，实在是太和谐、太安全了。而且，每八度只有5个音符仍然不算是很多，但是

如何添加更多的音符呢？这个问题引发了数百年的仔细实验和学术辩论，最终形成了一种被称为“十二平均律”的系统。在这个系统中，一个八度音阶被分成12个音符，相邻的两个音符之间相差一个半音，或者100音分。

在音乐中，重要的是音符的比例，而不是音符本身，因此半音和音分都是按比例设定的。每一个半音比前一个的频率高约6%。这意味着与半音对应的频率越高，音符就越高。

一些作曲家曾经尝试在每个八度音阶上使用超过12个音符。在从一个音符逐渐到另一个音符的过程中，额外的音符可能会短暂地出现(这被称为滑音，或滑奏)。然而，在所有的文化中，每个八度的“合适”音符几乎总是12个。

在定义了12个音符之后，人们开始在每个八度音阶中只使用其中的8个音符(因此得名八度音阶)。一个八度音阶中只有7个不同的音符——第八个和第一个是一样的，只是比第一个高一个八度(即使是能立即识别一个音符的专业音乐家也很难说出它是哪个八度)。

心理学家乔治·米勒(George Miller)在1956年解答了我们在一个八度之间只用12个音符中的8个的原因。他通过实验发现，我们的短时记忆最多只能同时存储7个项目，也就是只能有7个不同的音符。

整组12个音符中相邻的音符之间有一个半音的跳跃，所以我们选择的任何7个音符组成的音阶都会间隔一些全音(两个半音)。这样我们才能有ABCDEFG这样的音阶。音阶是每个音符都高于前一个音符的任意序列。

我们选择了7个不同的音符来定义音阶，但这并不意味着我们不能在一段音乐中使用其他音符。如果想要使用那些“间隙”音符，我们

可以用符号来标记它们，这些符号表示“比音阶中下一个最低音符高半音阶”或“比下一个最高音符低半音阶”。

在上面我们给出的音阶例子中，在第一个间隔中产生的额外音符被称为升A(升号记作♯)或者降B(降号记作♭)。如有需要，音符还可以进一步细分，这种更小的音阶被称为微分音。

在写一首乐曲时，通常有一个比其他音符更重要的音符——主调或主音。通常，这个音符是一首歌开始和结束的音符。在《 飞跃彩虹》 中，主音是落在Some和I上的。在影片中朱迪·加兰演唱时主音是降E(♭E)。因此，她唱的版本被称为降E调。

如何选定一个音阶呢？乐曲的情绪是主要的考虑因素。如果作曲家写的是一首自信、欢快、节奏快的曲子，他可能会选择大调。要写一首情感不太清晰的曲子，通常要选一个小调。大调和小调音阶的主要客观区别是(自然)大调音阶主要由主音(可以是任何音符)再加上6个与主音搭配最和谐的音符。“搭配最和谐”的意思是它们的波长比是整数的概率最大。因此，声波结合在一起形成了最规则的模式。在实践中，我们拿一个五声音阶并增加两个音[3]。这两个音符中的一个叫导音，其频率是音阶最低音的 。

在一个自然的小调音阶中是没有五声音阶中的两个音符以及导音的(它们被下移了一个半音)。其结果是，不再有大量的简单整数比，结果更模糊、更复杂，听起来更不完整。

此外，当我们听到任何演唱或演奏时，音符总是伴随着完整的和声(除了一些电子乐器)。大调音阶也包含所有这些和声，小调音阶则不包含。因此，在大调音阶中有一种真正的“自然性”或“完备性”，而这在小调音阶中是没有的。《 奇异恩典》 和《 我的女孩》 都属于大调，《 我射杀了警长》 属于小调。

过去(现在也仍然)有人声称，除了大调和小调的区别，每个调都有自己的情感特质。大量的实验证明事实并非如此。这一想法很可能是受著名或有影响力的音乐作品的影响而产生的——模仿往往具有相同的基调，因此与原作的情感品质有关。这并不是说曲调的选择是任意的，有些乐器的设计是为了让特定曲调从生理角度来说更容易演奏一些。

声音序列

虽然世界各地音乐创作的“原材料”相似，但在发展上却有很大的不同。在西方音乐中，使用的音符相对较少，但它们经常被同时演奏(和弦)或相互关联(和声)。如果我们同时处理7个以上不同的音符，作曲就会更加困难。特别是，同时演奏或演唱不同旋律的复调乐曲将变得几乎不可能。在东方音乐中，作曲家往往使用更多的音符，但更少和声。这种方法上的差异也通过乐器设计体现出来。西方的乐器，如吉他和管风琴，是复调的；而东方的乐器，如锡塔琴，通常一次只能发出一个音符，但可以很容易快速而微妙地改变那个音符。

西方音乐和其他音乐之间的差异可能起源于10世纪，那时复调音乐诞生于欧洲(尽管有些人认为复调音乐的诞生要早得多)。从复调音乐中产生了西方特有的多音调方法和发展这种方法所需要的和声研究。有人认为，第一次进行复调实验是为了让男人、女人和孩子们一起唱歌。

旋律

按顺序演奏一些音符，你就能得到一段曲调，从技术上讲就是一段旋律了。如果我们不考虑不同的音高，那么剩下的就只有所谓的节奏，即音符的相对长度和休止符，演奏整首曲子的速度(节奏)，以及

节奏的模式(节拍，或称拍子)。这不仅代表单词“韵律”这种技术意义，它其实是指音长的模式。在《 飞越彩虹》 中，从技术角度看它的节奏就是“长、长、中、短、短、短、中、中”。

节拍是由时间拍号标出的。最普通的节拍四四拍的拍号是$\frac{4}{4}$。分母上的数字给出了用来指定节拍的单位；分子指定了每组的节拍数。在乐谱中，每组的结尾都用一条竖线表示，这条线叫作小节线(四四拍指每小节有4个四分音符)。

音乐中的重要拍子类型并不只有这一种，两个波长相似的音调同时演奏时会产生另一种拍子。每隔一段时间，两种声波的波峰或波谷就会重合，叠加在一起，形成特高压区或特低压区。这种效应被称为低频“拍”，拍的频率等于两个声波频率之差(见图13)。

观看任何一部讲述伦敦警察的黑白惊悚片，当警察吹哨子时，你都会听到节拍。这种哨子用两根长度略有不同的管子来发出引人注意的声音。但节拍也并不总是令人不快，有一些管风琴中包含一个名为**天籁音栓**(voix c leste)的止音器，它控制着一组稍微走调的管子。当与普通的管子一起演奏时，产生的节拍会发出令人愉快的摇摆声。

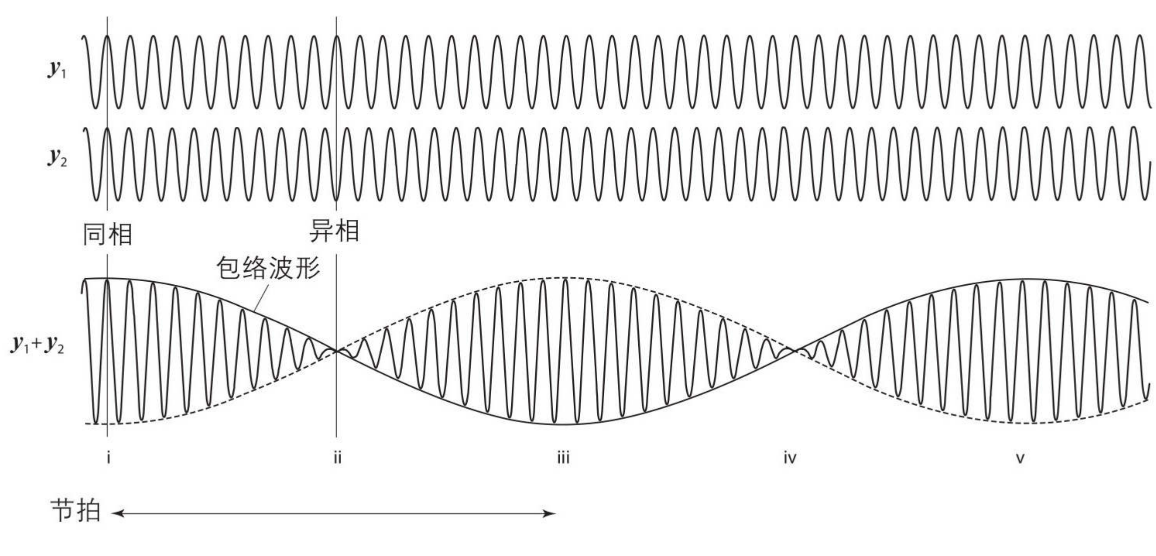


图13　节拍

间奏

传统的乐器是根据它们最初是由什么材料制作来决定分类的，因此有弦乐器、铜管乐器、木管乐器和其他乐器(比如打击乐器)。另一种方法是通过乐器是有音调的(像钢琴)还是无音调的(如沙球)来决定分类。任何音乐都需要节奏，使用没有音调的音源可以避免与“旋律”乐器产生不和谐。节奏源不一定是一种乐器，让-巴普提斯特·卢利(Jean-Baptiste Lully)曾用一根大棒敲打地面。这已经不流行了，部分原因可能是卢利死于这种演奏：大棒打到了脚背，导致他患上了晚期坏疽。鼓很快成为一个更受欢迎的音源，尽管它们也被用作“实际”的乐器。低音鼓和小军鼓这样的鼓是没有音调的，但定音鼓有音调。

大多数有音调的乐器都是通过敲击或拨弄弦、线、板或其他固体来发声，也有一些通过肺部或风箱提供能量使空气柱振动来发声。在这些情况下，乐器由共振加上泛音(在有铃铛的时候还有底音，也称为次谐波)来提供所需的声音。

这使得制作乐器看似挺简单，但实际上有三个难点。第一点，像金属板这样的物体不仅在两倍于其长度的波长上共振，在两倍于其宽度和厚度的波长上也共振，它旋转时还会发出其他声音。

对弦乐器演奏者来说，第二点非常有用。当一个物体，比如说120厘米长的弦被击打(或弹奏、运弓)，它的运动模式不仅取决于它的长度，还取决于击打的位置点。因为击弦点只能是一个波腹，而弦的末端只能是节点，所以如果弦在距离端点40厘米的地方被击打，它就会形成一个节点与反节点距离为40厘米的基频，因此波长会达到160厘米。此外，还会有一种节点与反节点距离为80厘米的声音。当弦的振

动变成它的固有振动模式时，这些成分会迅速消失，这种振动模式的中心有一个腹点，因此会产生240厘米波长的音调。主观上的影响是，近中点拨弦发出的声音更柔和，近端点拨弦发出的声音更刺耳。

风笛和管乐器的工作方式与此类似。在一个开放的管中，最常见的是有一个可以自由运动的开口端和不能自由运动的封闭端(封闭端不可能有自由运动)。12厘米长的管会有一个基本节点在0厘米处，第一波腹在12厘米(由于终端修正会多出一点点)，因此波长约48厘米。但是，就像在弦乐器中的情况一样，这还没完。虽然基频是由管的长度来定义的，但是其他的共振会根据管道的宽度来产生。这些与基本音及其泛音的混合产生了不同的音色，在更宽的管子里，较低谐波中有更多的能量，音色更圆润，而更薄的管子音色则更明亮(或更尖锐，每个人的感觉不同，因而描述也会不同)。

乐器制造商最后要考虑的是，一根位于两点之间的弦只会发出非常小的声音，因为它非常薄的横截面只会影响到周围很小体积的空气。为了把弦变成乐器，必须使更大体积的空气运动。在电吉他中，这是通过使用金属弦的振动在拾音器(用细线缠绕的磁铁)中产生小电流来实现的。在原声吉他和其他弦乐器中，顶部木质共鸣板受弦影响而振动，它的面积更大，比弦本身能够移动的空气更多。此外，在低频段，这些乐器的空心体充当亥姆霍兹谐振器。

非电子的乐器没有能量输入。相反，它们将琴弓、拨片或气流所提供的能量转化为一种声学形式。在某些情况下，当乐器的主体将频率移到4千赫左右时会听起来更响，尽管由于摩擦损失，声音功率实际上减少了一点。这种响度的增加源于我们听觉系统的结构，我们将在第4章中看到。

[1] 原文note统一译为音符。但实际原意更接近单音，和其让人产生的听觉感受。

[2] tempo是意大利语“乐曲速度”的意思。

[3]　指在五声音阶CDEGA中间添加F和B这两音构成七声调式，其中B就是那个导音。

04 听到声音 Hearing Sound

听力的范围

能够听到声音并不是一件稀奇的事情，因为强大的声音会振动身体，就连单细胞生物也能感觉到。但是，能够像人类一样听得那么清楚，简直就是奇迹。功率10～15瓦的声音移动鼓膜的幅度比氢原子的直径还小，即使是功率这么小的声音，我们也能够轻松听到。

我们能够听到的声音的强度范围也同样令人惊异。人类可以听到的声音中，最安静的声音(听力阈值，0分贝)和疼痛阈值(约130分贝)之间的差距十分巨大——130分贝，这个数值相当于价值1000亿英镑的便士数量(10^{13})。

我们还可以听到相当广泛的频率范围：大约10个八度，比钢琴键盘还多几个。而对于远远低于和高于这一频率范围的声音，虽然我们无法真正听到，但却可以感觉到，这一点将在第6章做出解释。我们对于声音的频率有非常好的辨别能力：大多数人都能察觉到大约1/4个半音的差异；在理想条件下，我们甚至能够分辨大约1/12个半音的差别。相比之下，我们对声音方向性的判断能力则相当一般。即使在十分有利的条件下，我们也只能确定水平方向10°左右、垂直方向20°左右的声源的方向。在这方面，许多其他动物都远远超过了我们。

也许我们所有听觉能力中最令人惊叹的一点是，即使谈话的声音不到背景噪声水平的10%(假设背景噪声的频率分布很广)，我们也能够听清楚：这远远超过任何机器的极限。

我们的耳朵有两种功能：听觉和平衡。平衡完全由半规管负责(见图14)。耳朵的其余部分就可以自由地根据我们的需求演化出最合适的

系统。但我们的发声器官却不是这样，作为人类进化过程中出现相对较晚的器官，它必须自己挤进脸颊和下巴之间，并与它们共享之前用于呼吸、进食、舔舐、吮吸、亲吻和战斗的那些身体结构。不过，对一个训练有素的演员或技艺娴熟的歌手而言，他们的发声系统可以像斯特拉迪瓦里小提琴一样完美地工作。

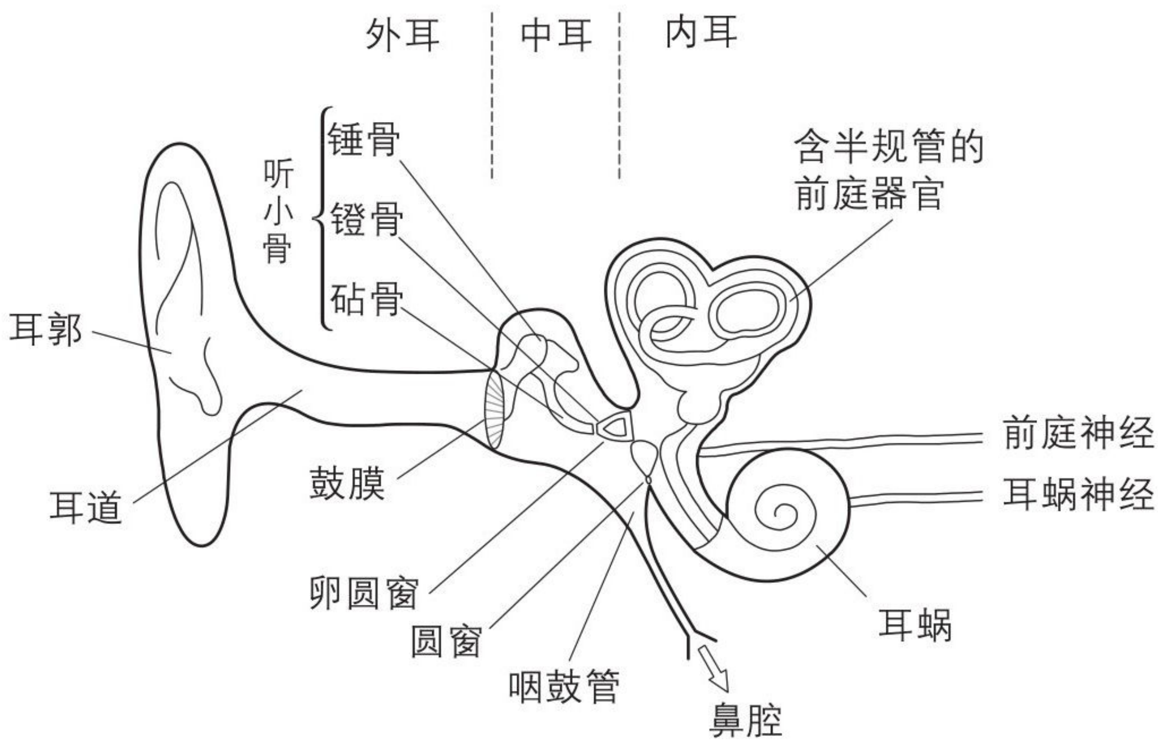


图14　耳朵(中耳和内耳经大比例放大)

耳朵能听到什么

如图15所示，我们的听觉系统还没有进化到可以测量声音的物理功率。例如：短笛的最大功率约为0.08瓦，而长号的最大功率约为6瓦，但长号听起来比短笛更安静，这意味着长号演奏者必须比短笛演奏者更加用力(实际上短笛演奏者从来没有在任何常见乐曲中全力演奏过)。实际上，所有乐器的功率都很低。一个响亮的管弦乐队全力演奏

的功率可能达到60瓦。如果这样的演奏持续两分钟，产生的能量足够煮开一汤匙的水。

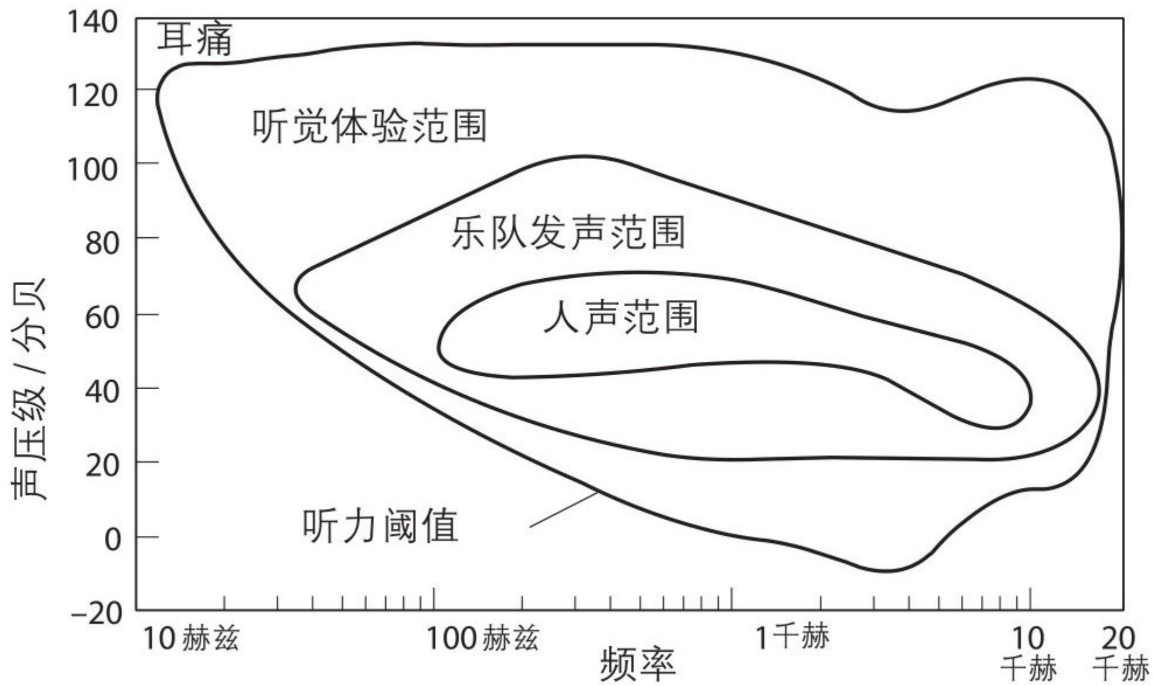


图15　听力与频率的相关性

此外，这些能量从声源产生后会向各个方向传播，因此，如果你坐在离这样一个管弦乐队10米远的地方，只有不到0.01%的能量会到达你的鼓膜。实际上，鼓膜所探测到的不过是空气带来的一系列快速撞击，除了它们撞击的力度和到达的速度之外，这一过程并没有传递其他任何信息。我们之所以能够体验到一个充满情感和富有意义的声音世界，要归功于解剖学、电化学和神经学上都高度进化的处理系统的精确协调。

外耳：捕捉声音

耳郭是耳朵的可见部分，像漏斗一样，主要起着收集声音的作用。耳郭前后不对称的形式也有助于收集声音在传播方向方面的信息。一些哺乳动物的外耳可以移动，有助于他们感知声音的方向。虽然有些人也有这种能力，但这除了娱乐价值之外没有任何用途。

由于耳道是一个约3厘米长的圆柱体，而共振波长为这个长度的两倍，那么相对应的频率为3000赫兹左右。该频率能量的增加会导致其他频率能量的损失，因此耳道就像一个带通滤波器。

在耳道的末端是**鼓膜**，一个约1厘米宽的圆形皮肤盘。它与耳道呈一定角度倾斜，以最大限度地扩大其面积，从而接收尽可能多的能量。鼓膜略呈圆锥形，这使得它比扁平形状能传递更多的能量。3000赫兹以上的频率在鼓膜上不会形成共振，鼓膜的表面以一种混沌的方式运动；而对较低的频率而言，由于其波长比鼓膜尺寸大，因此鼓膜会做整体运动。因此，它能够以极小的滤波形式传输尽可能广泛的频率范围。实现这种近乎平坦的频率响应一方面是通过外耳的不对称形状，另一方面则依靠胶原纤维的内部支架。

鼓膜必须紧绷着，但是又不能僵硬，这一点很重要，是通过咽鼓管对内外部压力进行平衡实现的(我们通常将咽鼓管画成开放的，但通常情况下它并不是开放的，除非有明显的压力变化导致其打开，这时会产生一种独特的咔咔声)。

中耳：声音增强

鼓膜与三块被称为听小骨的微小(实际上是最小的)骨骼相连，这些骨骼占据着充满空气的中耳。它们的主要作用是将鼓膜宽而浅的运动转化为对第二层鼓膜的高压“轻敲”，这层膜叫作“圆窗”，也是进入内耳的通道。听小骨起了杠杆的作用，使力增加了1.5倍。然而，

这种力增强的主要方式其实是通过鼓膜和圆窗的面积之比实现的，将力集中在一个更小的区域，从而将压力增加大约20倍。听小骨还通过听觉反射为内耳提供一些保护(见第8章)。

内耳：从声音到电流

内耳充满液体，就像鼓膜把通过空气传播的声音转化为通过骨头传播的声音一样，圆窗可以将骨骼传播的声音转化成通过液体传播的声音。声音将沿着一个被称为耳蜗的长约2厘米的盘管传播。在它的末端有一个孔(螺旋孔)，声波通过这个孔后沿着第二根管道传播回来，第二根管道沿着第一根管道的延伸长度与其连接。当声波完成它的双程传播返回时，它必须被消除，否则它会反射回耳蜗干扰新到达的声波。所以，第二管终止于另一层膜——卵圆窗。当声波到达卵圆窗时，卵圆窗向外膨胀，将声能转化成热能，然后消散。

基底膜将声波转换成神经脉冲，它像长棍面包的馅料一样在耳蜗的管子之间流动。如果没有像蜗牛那样卷起来(拉丁语中，耳蜗即蜗牛)，耳蜗中并排的管子就会长达5厘米。由于耳蜗的长度与声波的波长有关，因此所有哺乳动物的耳蜗长度都非常类似：大象的耳蜗也只比人类的长50%，所以卷曲可能只是一个节省空间的方式，不具有任何声学功能。小鼠和其他小型哺乳动物无法容纳全尺寸的耳蜗。它们的耳蜗长约1厘米，因此它们只能听到3～4个八度的音程，相比之下，我们和大多数其他大型动物都可以听到8～10个八度的音程。在基底膜上的是柯蒂氏器，其上生长了9排短毛(每排约400个)，称为**静纤毛**(stereocilia)。这些短毛沿着膜的长度延伸，并且有神经纤维与它们相连。这些纤维聚集在一起形成听觉(耳蜗)神经，其将脉冲传递到大脑。

基底膜随着声波的敲击而运动。它的底部比尖端更硬更宽，这意味着较低频率的声音会引起靠近尖端部分的振荡。这些振荡导致静纤毛移动，然后附着有静纤毛的毛细胞向大脑发送电化学脉冲。由于大脑知道各个毛细胞的位置，它可以通过这种方式确定声音的频率(这被称为听觉系统的位置说)。然而，当声音低于1千赫时，整个基底膜都会振荡。在这种情况下，一些其他的机制变得重要起来：静纤毛细胞随着声波脉冲发射信号。举例来说，对一个100赫兹的音调，静纤毛细胞每秒发射100次信号。

然而，细胞无法在一秒内发射超过500次信号。因此，为了应对更高的频率，它们面临的问题与一支装备了燧发枪的军队遇到的问题类似，如果说重新装弹需要一分钟，那么小队怎么能产生间隔10秒的持续的枪声呢？答案是将队员分成6组。第一组发射并开始装弹，10秒后第二组发射，以此类推。在第六组开火10秒后，第一组将完成装弹并将再次开火。毛细胞正是以这种方式分组工作的。假设第一组在声波周期达到峰值时向大脑发出信号，第二组则在该周期已经下降到最小值的一半时发出信号，第三组在周期最小值时发出信号，而第四组则在周期返回至最大值的一半时发出信号。以这种方式，这样的四个单元组可以对频率比其最大发射频率高4倍的音调进行响应。

基底膜上的毛细胞也以一种非常不同的方式协同工作。那些排在基底膜第八排或者更外层的毛细胞通过瞬时改变它们的长度来响应传入的声波。这种运动放大了内层毛细胞(唯一向大脑发送信号的毛细胞)上的立体纤毛的振动，因此产生了显著的放大效果。这样我们能够听到的声音的最低值又降低了40分贝(到了我们原本听觉下限的万分之一)。这种活动会产生微弱的声音，称为耳声发射。因为婴儿太小，无法说清他们是否听到或到底听到了什么，所以耳声发射对于确定婴儿听力系统的功能非常有用。此外，当内耳受到损伤时，耳声发射就会减弱，所以它们也是听力学专家检查听力情况的有效手段。这种声音

太安静了(是件幸事)，人基本上无法听见，所以要用非常灵敏的入耳麦克风进行测量。

神经与大脑：从客观到主观

从基底膜发出的神经信号并不是模仿声波，而是包含三个信息的编码消息，即：(a)有多少神经纤维在同时发射信号；(b)这些纤维沿着基底膜排了多远；(c)纤维信号发射的时间间隔是多久。大脑通过(a)和(c)可以提取响度信息，通过(b)和(c)可以提取音高信息。

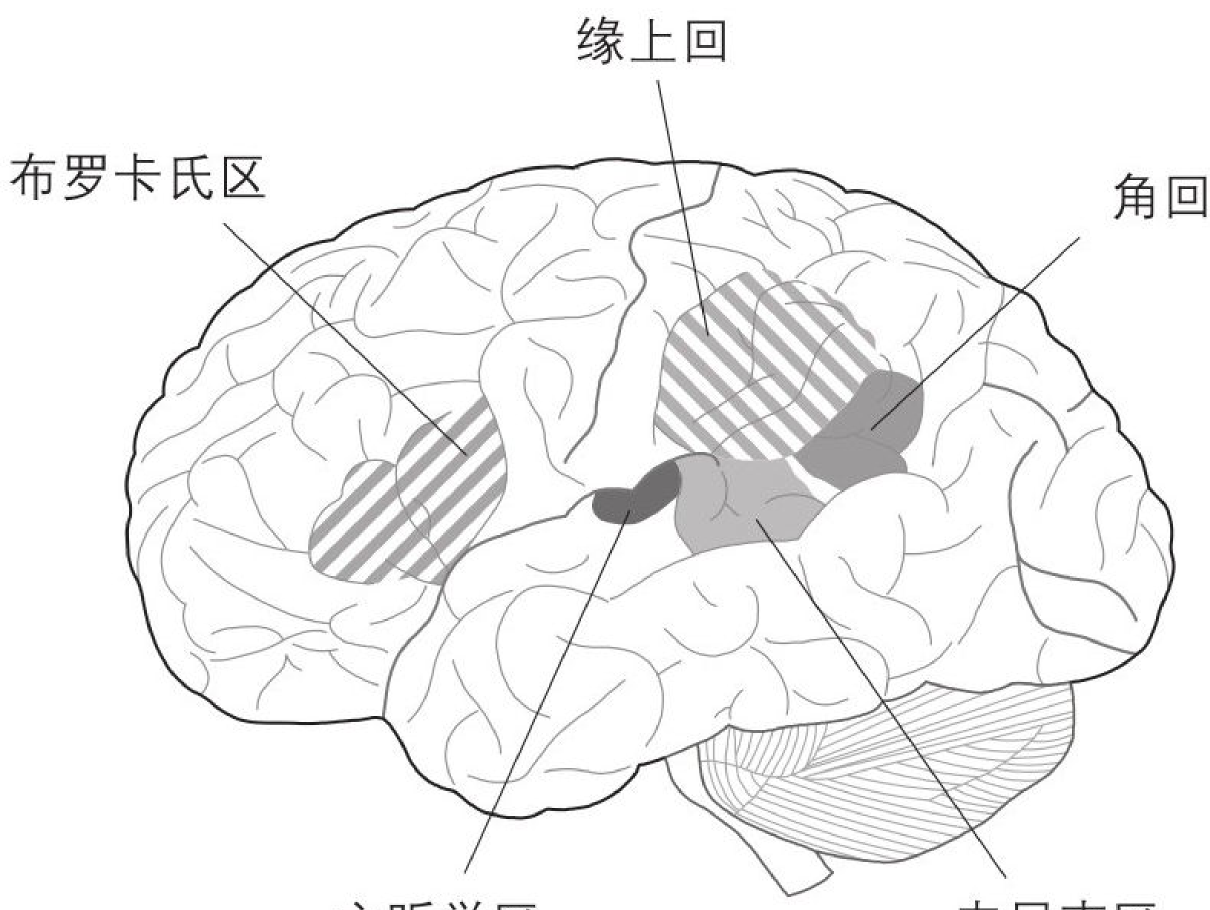


图16　大脑中听觉、语言和言语活动的位置。主听觉区识别声音，布罗卡氏区既分析又生成语义元素，韦尼克区处理的是语音序列(听到的、发出的和记住的声音)，

缘上回负责发音，角回负责语义处理

我们还没有完全弄清在大脑的听觉和语言中枢(见图16)里发生了什么，但处理过程的第一阶段是从听觉神经的输入数据流中提取显著特征。这些特征被用来不断地更新、修改和精炼我们脑海中关于所听到的事物的模型，比如一个曲调、一句话，或者一段令人烦躁的引擎噪声。大脑通过预测声音接下来会怎么变化来测试模型的准确性。在预测的过程中，大脑会尝试确定声音的每个组成部分对于这段声音的含义所做出的贡献大小，这个过程被称为层次编码。例如，如果有人试图哼出《 飞越彩虹》，哪怕他哼错了大部分调调，我们也能识别出这首歌，这也证明了我们在声音预测方面有惊人的能力。哺乳动物的大脑对传到耳朵里的声音的理解速度几乎令人难以置信——对一只狗说“松鼠 ”，如果它能听懂这个词，它几乎能瞬间对其做出反应。

回归

从进化的角度来说，听音乐对大脑来说是一件相当不寻常且新奇的事情。通常在听音乐的过程中，兴趣的焦点是外部世界的某个物体，因此听觉系统的一个重要功能就是定位那个物体。对于高频的声音，来自左侧的声音将首先到达左耳，同时由于头部对声音的阻挡，大脑会感觉到右耳处声音的音量比左耳低。当声波的长度大于两耳之间的距离时，大脑会比较两耳中声音的变化程度。如果一个长波从左侧传来，它的每一个波腹都会率先到达左耳，所以左耳的压力刚开始是最大的。随着波的传播，左耳的压力下降，而右耳的压力上升，直到波腹通过，压力再次下降。然而，当一个波的波长超过4米时，从左耳到右耳的这段距离不会有什么变化，因此无法判断它的方向。

大脑用头部和肩部形状对声音的影响来确定声源是在双耳水平的上方还是下方。拥有两只耳朵的优势不仅是可以判定声音的方向，还

有其他用途，例如，有的时候即使外界没有声音，听觉神经也会发出信号，但如果这些信号只来自一侧，大脑就会拒绝接收。

大脑的处理系统已经进化到会对它接收到的声音做出合理的假设，这导致了诸如优先效应(也被称为第一波前定律或哈斯效应)这样的现象。大脑的假设是，第一秒到达的声音表明了声源的方向。因此，随后的声音被认为与第一个声音来自同一个方向。这使我们能够在黑暗的空间中定位声源，而不会被来自四面八方的回声所混淆。但这样的假设会产生误导，尤其是在那些非自然发生的情况下。比如，让接收者听一个位于偏左45°方向、距离大约1米远的扬声器传来的声音，然后用另外一个更加响亮且完全相同的声音将其逐渐取代，只是这个新的声音来自右侧45°方向的扬声器，但是接收者听到的声音似乎仍来自左边。

对试图在人群中交流的人来说，鸡尾酒会效应会非常有用。在鸡尾酒会效应中，特定的词句(比如一个人的名字)会从嘈杂声中赫然而出。这也适用于非言语的声音：指挥家往往对特定的乐器或乐句高度敏感。这种效果之所以产生，是因为无论我们是否在主动倾听，大脑都在不断地建立模型，优先找出有重要意义的声音，比如人的名字。

听觉是非常丰富且复杂的，正如语言学家罗兰·巴特(Roland Barthes)所指出的那样，声音以三种方式作用于我们的大脑：作为“指示符号”(令人震惊的爆炸声)、作为“符号”(一个单词的字面意思)和作为“能指”(由“结束”这样的单词引发的无意识联想)。根据拉贝尔的说法，听力在很大程度上也是一种社交活动：

> 丰富且有起伏的听觉内容在很大程度上打破了私人和公众之间的界限。声音起作用是通过形成连接、群组和结合，这强调作为关系投射的个人身份……把一个人编织进一个更大的社会结构中……为实现共享空间的意义做出贡献。

拉贝尔还指出，无论我们喜欢与否，声音都让我们与他人——哭闹的婴儿、吵闹的邻居或欢呼的足球支持者产生亲密接触。正如他所说："声音创造了一种关系地图，这种关系往往是情绪化的、有争议的、流动的。"

听小骨

鼓膜是造化"工程"的奇迹，但实际上没有鼓膜我们也能听清，因为声波也可以通过头部的骨头——特别是耳后的乳突骨到达内耳。将耳朵没入浴缸，在很大程度上等于切断了空气传播的途径，所以这时声音主要是通过骨骼传播到达大脑的。然而，乳突骨传导系统相当不敏感。使用空气传导，我们可以听到的声音比乳突骨所能检测到的最弱声音还低40分贝左右。但是，另一方面，通过骨传导我们能够听到频率高至30千赫的声音，这远高于空气传导的最大频率。不过，可能过高频率的声音对我们来说没什么价值，它们都以同样的方式编码，那些高频的声音和20千赫的声音听起来音高是一样的。

宽吻海豚的骨骼听觉系统要先进得多。它们的下颚长着牙齿，牙齿间隔有规律，角度相同，形状非常相似，且各自高度取决于它们的位置。这些加起来就构成了一个**聚焦阵列**，只要声源位于正前方，特定波长的声波就会被显著放大。因此，海豚可以听到非常小的声音，并且可以通过把头转动到响度最大的方向来定位声音。

耳聋

听力系统是一个脆弱的系统，耳膜或听骨受到严重损害的情况并不罕见。当这种情况发生时，我们必须依靠骨传导以及人工辅助。例如，爱迪生用他的牙齿把声音从留声机传到乳突骨，那台留声机上现在仍能看见牙印，述说着当年的故事。

上面这种情况称为传导性听力损失。如果内耳或听神经受损，其结果是感音神经性(或说“神经性”)听力损失，影响的主要是人对高频率低强度声音的听力。在温和的情况下，它会导致一种叫作“响度重振”的情况，在这种情况下，声音的“可听性”会出现突变。在一个声音响度逐渐增加的环境中，有的人一开始什么也听不见，之后会突然听到特别响亮的声音。因此，“没有必要大喊大叫”的抗议活动就是针对那些为了让对方能听到自己的声音而突然提高嗓门说话的人。

感音神经性听力损失是最常见的类型，其最常见的原因是毛细胞受到物理损伤。当声音非常高时，鼓膜会破裂(鼓膜也会因头部受到撞击或感染而受损)。然而，值得庆幸的是，鼓膜破裂不仅可以愈合，而且通常能恢复得和之前一样好。

全世界大约有3.6亿人(超过全球人口的5%)患有“严重妨碍健康”的听力损失。相对于听力较好的人，患病成年人的听力损失大于40分贝，患病儿童的听力损失大于30分贝(儿童约占患者总数的10%)，则可以定义为患有严重妨碍健康的听力损失。年龄在65岁以上的人中大约有1/3正遭受着这种听力损失的折磨。

在治疗耳聋方面，根据每位患者的具体情况，可以采取不同的治疗手段。对于传导性听力损失，如果患者还保有部分听力功能，根据佩戴者的特定听力损失模式量身定制助听器是最有效的办法。这种助听器通常也具备降噪功能，并且会使用定向传声器，以便患者能够集中于他正看着的声源。这类助听器还可以加装一个振动元件来刺激乳突骨。助听器在治疗感音神经性听力损失方面不太成功，因为信号会被响度重振效应所扭曲。

对完全失聪的人来说，耳聋治疗困难更大，但在过去几年中，人工耳蜗植入的出现给他们带来了新希望。2012年，通过刺激干细胞生

长成毛细胞，失聪的沙鼠能够恢复45%的听力，少数沙鼠甚至恢复了90%的听力。未来的某天，这种方法可能会适用于那些因螺旋神经节神经元受损而丧失听力的人(这类人约占失聪人群的10%)。一些动物失去静纤毛后很容易就能再长出来，如猫头鹰，或许可以通过基因诱导，将这种功能用来造福人类。但猫头鹰听力中其他独有的特性我们并不会使用，比如，猫头鹰的听力在春季时最佳，这很可能是因为春季它们需要为自己的雏鸟捕捉更多的猎物。

除了听力损失，还有其他很多与听力相关的问题。最常见的就是耳鸣，也就是耳朵里嗡嗡响。其成因还不太清楚，而且声音的响度、类型和持续时间很大程度上会因人而异。它通常与过去的感染、用药(特别是某些抗生素)或创伤有关，而且常常伴随着听力损失。

语言的结构

大约100万年前，我们祖先的听觉系统产生了微妙的变化，从而能够分辨语言。人类对自己发声系统的进化知之甚少。虽然大多数动物都有以可控的方式进行发声的能力，但相较而言，人类语言的复杂性却是无法估量的。因此，与研究其他器官(比如腿)的进化不同，我们不能回顾过去那些古老形态，研究其种种发展形势，然后探究人类怎样适应这些变化，并进化出与之相配的系统。

在发声系统最基本的形式中，发出声音非常简单。图17显示了发声所涉及的结构。空气通过一根管子(气管)从肺部排出，管子上装有片状的发声褶皱(声带)，它限制空气流动，当肌肉绷紧时就会振动。增加这种张力会增加振动频率，但是褶皱的长度会为频率设置下限，这导致男性声音的基频为125赫兹，女性为200赫兹，儿童则为300赫兹(或更高)。对男孩来说，青春期时的声带长度会突然增加，这就导致了声音的变化，即我们通常所说的变声。

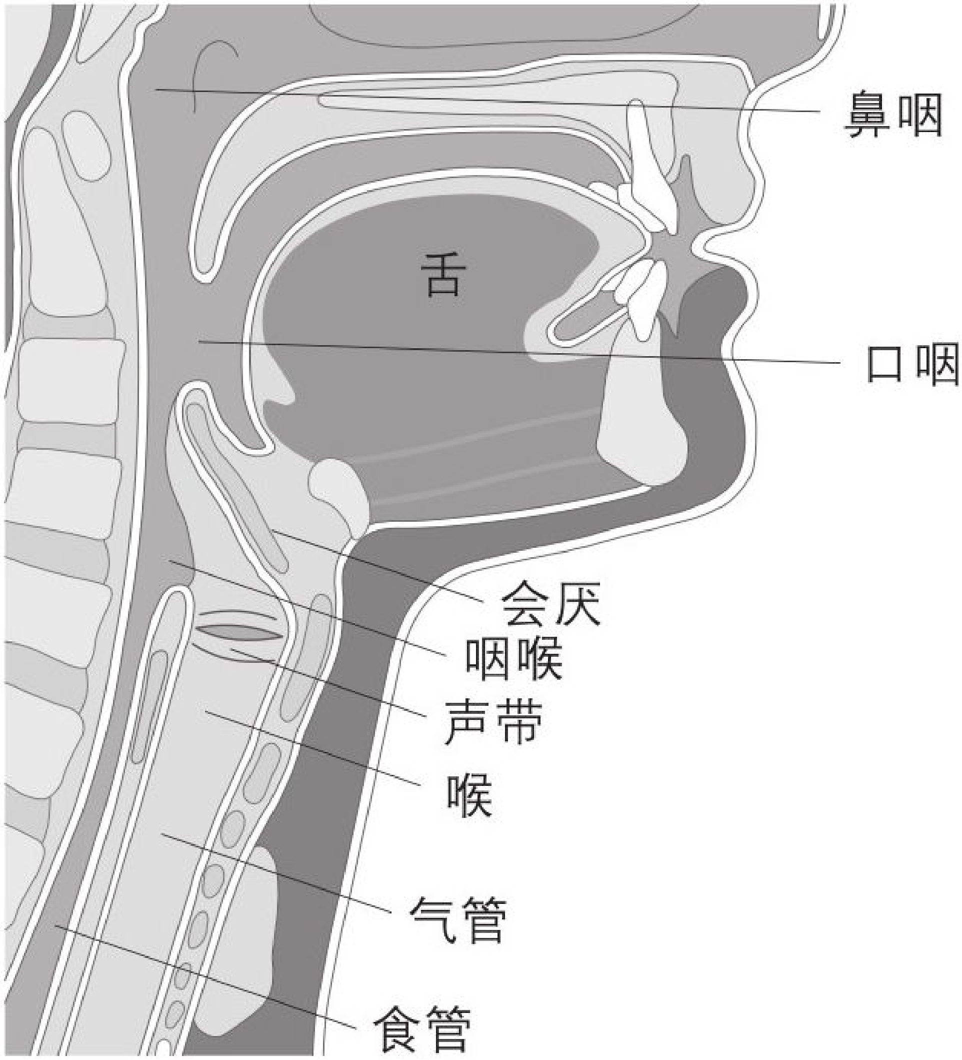


图17　发音器官

低频声音从气管末端发出后，进入由软腭覆盖的声道后部，在声道的前面是硬腭，这样形成了一个腔，其中的声音形成名为**共振峰**(formant)的共振。元音的特征波长就在这里被设置好，然后通过抬高舌头来改变声道的体积，或将声道分成两个相连的腔来改变元音的波长。

辅音比元音涉及更多的发声器官，通常持续时间较短，而且在许多情况下，辅音在发声过程中就会发生变化。辅音主要有四种类型，

分别由它们的发音方式来定义。

爆破音是因气流突然停止而形成的(因此它们的另一个名字是塞音)。摩擦音和流音则分别需要带有和不带有湍流的部分塞音。鼻音把气流转向鼻腔。滑音(半元音)包括从一个元音到另一个元音的快速过渡。如表4所示，这个命名系统的不同寻常之处是可以根据发音的位置对辅音进行细分。它同时也显示了辅音是否发声，也就是说声带是否参与发声。(腹语演员试图发出唇音或唇齿音的这一过程会出现需要抵抗嘴唇不自主开合的障碍。有技巧的腹语艺术家可以通过发这些音时加快语速，来规避这个问题。)

由于声道共振的波长只取决于其结构，所以改变声速就会改变这些波长的频率。因此，吸入氦气(密度为空气的14%)会发出“唐老鸭”的声音，而吸入氙气(密度是空气的4.6倍)会发出更为罕见的沙哑声。

发音方式		发音位置						
		唇音（唇）	唇齿音（唇和牙）	齿音（牙）	齿龈音（牙龈）	腭音（硬腭）	软腭音（软腭）	声门
爆破音	发声	b			d		g	
	不发声	p			t		k	
摩擦音	发声		v	th(en)	z	zh		
	不发声		f	th(in)	s	sh		h
鼻音	发声	m			n		ng	
滑音		w			y			
流音					l和r			

表4　英语辅音

然而，如果我们只是用我们的发声器官说话，生活将是一件枯燥的事情。从生理学上讲，唱歌和说话没有什么不同，只是发声的每一个方面都受到了更精细的控制，音高常常与一个额外定义的值关联。吹口哨则不涉及声带，它需要在嘴唇周围产生湍流，将能量转移到声腔，声腔起着亥姆霍兹谐振器的作用。要大喊大叫，只需从肺部获得更大的空气力量即可。在窃窃私语中，发声器官的工作原理与正常说话时一样，只是这时声带既没有振动，也没有完全放松，因此当空气通过它们时，会产生湍流(这被称为内收)。由于更多的空气可以在声带之间通过，而不激发声波，所以低语必然是相对安静的。

我们的听觉系统比我们最先进的机器都要复杂得多，并且已经进化得非常适合我们，但大自然赋予我们的能力是有限的。除极少部分人外，我们现今所使用的通信技术中的声音，已经大大超过大部分人的发声或收听范围。正是为了满足我们对通信的需求，我们才利用电流，甚至是电子，来满足这一需求。第5章将围绕如何实现这个功能而展开。

05 电子声 Electronic Sound

从声音到电流：麦克风

19世纪20年代，查尔斯·惠斯通(Charles Wheatstone)发明了麦克风。它是一种由两块金属板组成的纯声学设备，通过一根有弹性的杆子和一段女装裁缝常用的丝带固定在耳朵上。使用的时候需要把自己的头靠在声源上(惠斯通很贴心地建议使用一个烧水壶)，如果运气好的话就可以听得更清楚。意料之中(惠斯通却觉得很意外)，这个设备并没有流行起来。

现今声学设备的前身是碳质麦克风，由大卫·休斯(David Hughes)于19世纪70年代发明。与惠斯通的设备不同，碳质麦克风可以将声音转化成电。在这种设备中，一个薄金属片(膜片)压住充满碳颗粒的容器，下压的力改变碳颗粒的电阻，电流从中流过。虽然性能很差，但碳质麦克风在电话中使用了几十年。

虽然在过去的几年里，出现了很多种形式的麦克风设计，几种专业的类型现在还有，但目前常用的只有三种类型，即动圈(或动态)麦克风、电容式麦克风和压电式麦克风。

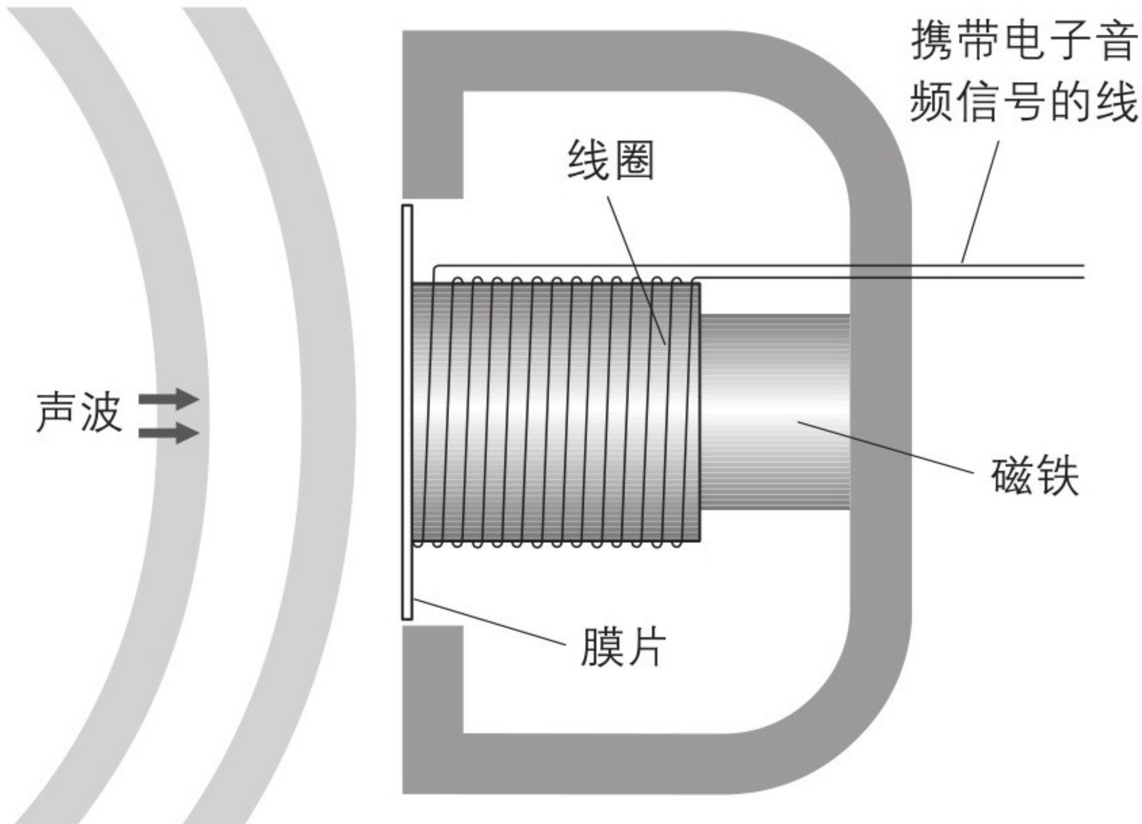


图18　动圈麦克风

在动圈麦克风(见图18)中，膜片连接到环绕固定磁铁的线圈上。线圈运动时产生低电压，从而产生一股弱电流。低电压意味着这种麦克风的质量不足以进行测量工作，所以动圈麦克风主要用在音乐会和录音棚中。

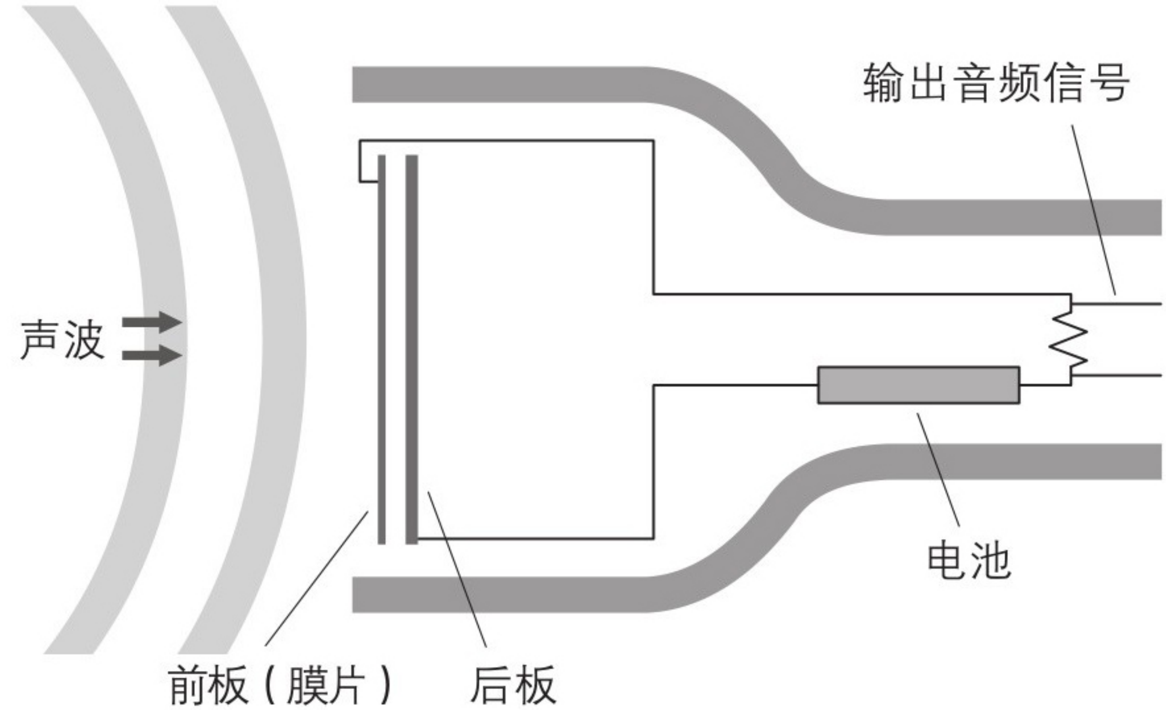


图19　电容式麦克风

在电容式麦克风(见图19)中，膜片形成一个电容板(因此得名)。电容器由一对平行的金属板组成，其间有一层薄薄的空气或其他不导电的材料(称为电介质)。将一块金属板连接到电池的负极上，电池就会给这块板充电。金属都含有自由电子，而另一块金属板上的自由电子会受到和负极相连的金属板上电场的排斥。被排斥的电子从平板中流出，从而使金属板带正电荷。因此，现在整个电容器上都有一个电压(称为极化电压)，麦克风可以随时工作了。膜片板随着声波的压缩和膨胀而移动。巨大的电阻阻止电荷快速逸出，因此声波被转换成电压波动的模式。电容式麦克风具有良好的频率响应特性，在实验室和声级计中被用作测量麦克风。它们对突发声音的反应也比动圈麦克风快。

晶体麦克风和陶瓷麦克风利用了压电效应，这种效应是利用了石英或其他一些晶体材料在受到轻微压缩时会产生电压这一性质。大多

数固定电话都使用这种麦克风，呼叫中心使用的受话器里也是它。

出于我个人对历史或专业上的兴趣，我将其他类型的一些麦克风列在下面。

驻极体麦克风是由永久充电材料(在电学中相当于磁铁)制成的。它们的功能与电容式麦克风类似。

微电子机械系统(MEMS)麦克风是直接蚀刻在硅片上的电容式麦克风，面积只有几平方毫米，而且价格低廉，特别坚固，这意味着它们可以被用于手机及许多其他设备中。

光学麦克风使用闪亮的硅薄膜作为膜片，它可以反射来自发光二极管(LED)的光。当薄膜被声波振动时，光探测器测量光的变化，然后电子电路将这些变化转换成电信号。这种麦克风结构紧凑、坚固耐用，不受局部电磁场的影响，因此，在病人进行核磁共振扫描期间，他们可以用这种麦克风和工作人员进行交流。

压力区麦克风是专门设计在硬反射表面附近使用的(如必须直接朝向舞台地板放置)。对于传统的麦克风，麦克风表面的反射会干扰麦克风直接接收到的声音，但是压力区麦克风克服了这个问题，因为它的膜片离麦克风表面非常近，以至于大多数波长都会重叠。

带式麦克风使用金属带而不是小的膜片，与鼓膜和大多数其他类型的麦克风不同，它对金属带两侧的压力差产生回应。这意味着金属检测到的是随声波移动的空气分子的速度，而不是声压模式。通常解说员会在嘈杂的环境中使用这样的麦克风，因为来自四面八方的声音会同时影响金属带的两侧，从而使得麦克风不会对这些声音做出反应。在这个应用场景中，麦克风在靠近上唇的地方有个凸起，这有助于引导解说员的声音，使它只撞击金属带的一侧。

在嘈杂的环境中准确捕捉声音的另一种选择是**项链式麦克风**，这是一种小型的可以夹在物体上的驻极体麦克风或动圈麦克风。项链式麦克风的优点是体积小，要将它藏在衣服下面简直轻而易举。

声强探头由一对面对面的麦克风组成，它测量同一声波不同部分的压力，由此可以计算出分子速度并进一步得到声强。

选择麦克风的一个重要标准是方向性。比如，理想的全向麦克风对来自任何方向的声音都有着同样的敏感度。再比如，在捕获完整的音效场景时常常需要使用全向麦克风。单向麦克风只能从一个方向接收声音，非常适合在嘈杂的环境中接收语音或歌曲。

除带式麦克风和声强探头外，只要声波波长比膜片的尺寸大，那么对这种类型的声波麦克风都是全方向的。如要使麦克风主要针对从正前方或正后方(双向)传来的声音敏感，只需让麦克风前后方都向空气敞开。膜片不会对四面传来的声波做出太大的反应，因为前部压力的升降与后部压力的升降非常相似。但是任何入射到麦克风正面或背面的波都会很容易地被接收到。

实现定向的一个简单方法是将麦克风安装在截面为抛物线的反射器的焦点上。只要麦克风比反射器小，这个抛物面的形状就可以将入射声波反射到膜片上。另一种选择是将麦克风安装在一个管道的末端，管道两侧有裂缝，以制造**枪式麦克风**(又叫远程拾音器)。声波沿着管道畅通无阻地传播，一直传到麦克风，但来自其他方向的声波则通过缝隙进入麦克风。每一个这样的声音都会进入多个狭缝，形成许多版本，每个版本都有不同的相位。因此，这两种版本将在很大程度上通过相消干涉互相抵消。**霰弹式麦克风**广泛用于户外录音，常常配合摄像机一起使用。不过，这种麦克风对频率的依赖度很高。

从电到声：扬声器

扬声器与麦克风的原理恰恰相反。如果给一个动圈、电容、晶体或陶瓷麦克风提供一个变化的电流，这些麦克风将因振动产生声波(这样的麦克风因此被称为**互易换能器**)。大多数实际的扬声器是反过来的动圈麦克风，称为动圈式扬声器(见图20)。顾名思义，一个电信号被传送到一个线圈(音圈)，该线圈连接到一个锥形的膜片上。线圈包围着磁铁，信号在线圈形成的电磁场中使线圈和膜片发生运动。

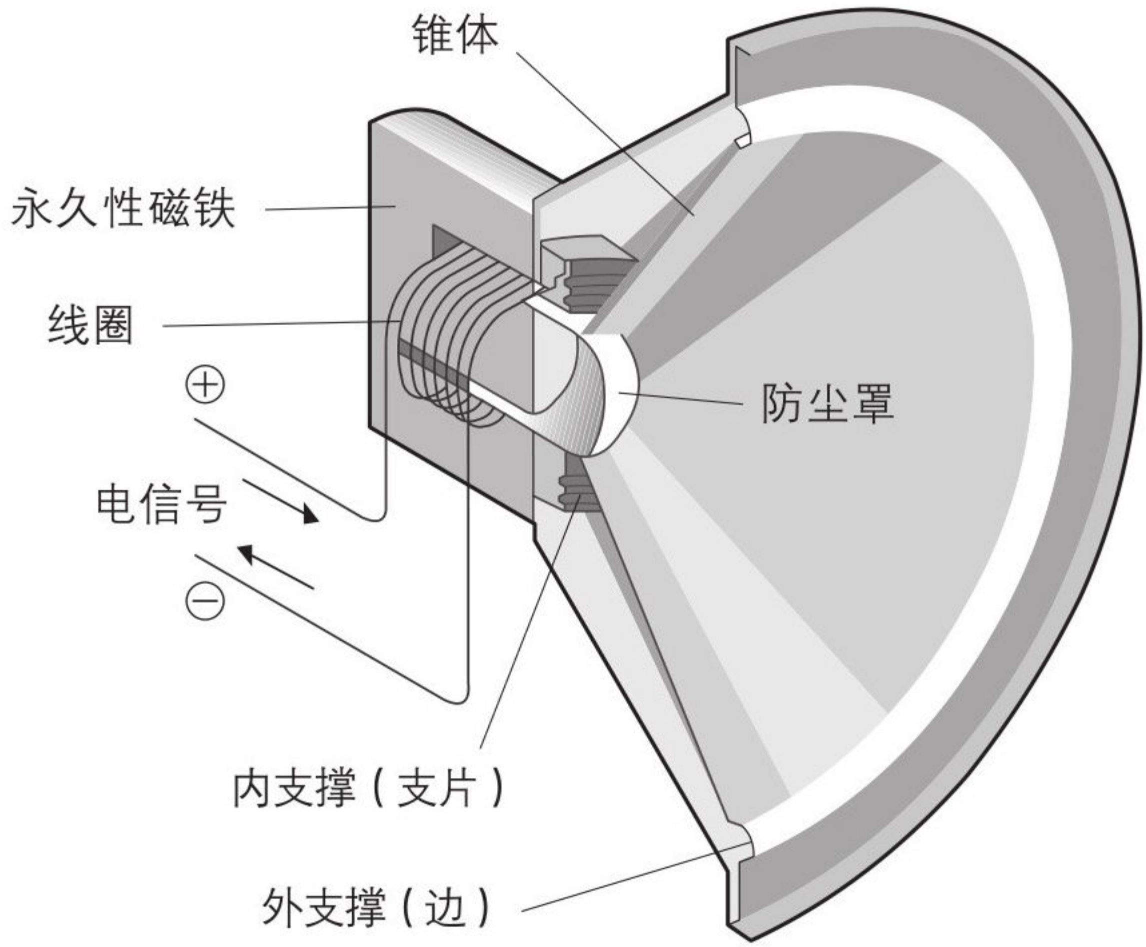


图20　扬声器

因为扬声器对电能的利用效率比较低，大部分进入扬声器的电能都转换成了热能，只有大约1%的能量可以被听到。因此，放大是至关重要的。如今，有了晶体管，放大变成了一件很简单的事情，主要挑战变成了确保每个频率被放大到一个适当的程度。考虑到听觉系统的非线性属性，如果要保持输出声音音高不变，就必须对不同的频率进行不同程度的放大。谨慎地选择放大倍数非常重要。除了会对听力产生有害影响，扬声器也很容易被损坏，尤其是人为制造的声音，因为这种声音可能会在一个极短的时间内快速变大。此外，如果麦克风发

送到扬声器的声音又反过来被麦克风接收到了，就会建立一个正反馈回路，导致一种我们非常熟悉的呼啸音产生。

扬声器是音乐生产链中最不**高保真**(hi-fi)的环节。虽然原理简单，但它们的设计会面临许多实际的挑战。扬声器必须保证当信号降至零时，音圈能精确地回到它的起始位置，而且在任何情况下都不会发生振荡，但是又必须处在可以自由移动的状态。锥体必须在振动时保持其形状不变，必须非常轻，但又足够坚硬，以免在重力作用下下垂。与此同时，它必须大到在低频时也能移动大量空气(足够强大的声波才能被听到)，但又小到足以在高频时每秒来回移动超过10 000次。此外，其外壳不能与任何频率产生共振。

实际上，分组使用扬声器要容易得多。这些扬声器通常都会组成一个单元，包括一个频率在2000赫兹以上的小型高音喇叭，一个较大的中频扬声器(50～5000赫兹)和一个低音扬声器(30～800赫兹)。对于那些喜欢低音的人，可能还会加一个低音炮(20～200赫兹)。

低音炮通常是有源扬声器，也就意味着它们有自己的放大器(因此需要配供一个电源)。其他大多数的扬声器都是无源的，由外在的信号进行驱动，这些信号常常通过高保真(或其他)系统中的放大器进行了增强。

一个没有外壳的扬声器几乎是无声的，原因很简单，扬声器前方产生的高压脉冲会悄无声息地溜到后方，填充刚刚在那里形成的低压区。因此，可以把扬声器放在一个密闭的盒子里。然而，如果盒子很小，当隔膜向外运动时，盒子里的空气很难被压缩。还有一种解决办法是将隔膜放在一个称为挡板的环的中心。挡板必须足够大，这样当压力脉冲跑到隔膜后面的时候，那里的低压区就已经消失了(换句话说，声波走的距离比最长波波长的1/4要长)。

亥姆霍兹谐振可以用来扩展扬声器在低频声方面的表现性能。在安装扬声器的盒子前面开一个孔，腔体就会与低频声音产生共振。如果盒子的共振频率低于扬声器的共振频率，当扬声器的膜片向初始位置移动(反射回波)时就会产生压力脉冲，脉冲在盒子中运动直到出现在盒子开口处，在那里它将与一个隔膜前方刚刚产生的新脉冲处于同相位。这两个同相脉冲会相互加强。这适用于一切脉冲，只要这些脉冲产生的声波频率在膜片和盒子的共振频率之间。

这样做的缺点是有时可以听到从端口吹来的气流，另一个缺点是声音不会那么清脆，因为每个信号后面都有一个短的衰减共振“尾巴”。此外，形成频率低于该端口频率的声波的脉冲会被来自膜片的后续波抵消。超过这两个共振频率的声波既不被增强也不被减弱。

我们的大脑非常善于填补声音的空白，我们可以利用这一点来改善基频非常有限的扬声器。在自然界中，一组200赫兹、300赫兹、400赫兹和500赫兹的音调几乎总是基频为100赫兹的谐波(泛音)。因为大脑的听觉中枢(见图16)“知道”这一点，所以它会自信地断定100赫兹确实存在。但是，如果一组音调来自一个小型扬声器，那么很可能实际上不会有100赫兹的音调。这种效应被称为缺失的基频，这也是为什么在埃皮达鲁斯剧院(见第1章)听那些缺失低频的音乐却不会觉得奇怪，因为观众的听觉中心自动补足了这些频率。这也解释了为什么尽管早期电话的低频传输信号并不好，

但是仍然有相对足够的清晰度。

高性能麦克风、扩音器和扬声器使唱片业发生了革命性的变化，而且由于唱片本身不需要更改，这种变革可以快速进行。紧接着，社会也发生了变化，一时间，几乎所有人都能接触到音乐，并选择自己喜欢的音乐。

声音研究专家乔纳森·斯特恩(Jonathan Sterne)称，这种新的声音媒体“对经验和存在的基础提出了质疑”。此外，这种声音媒体对表演者同样产生了深远的影响，正如音乐历史学家罗伯特·菲利普(Robert Philip)所指出的，许多人对他们在录制表演中听到的错误数量感到震惊。音乐学家马克·卡茨(Mark Katz)认为，这些表演者陷入了一个“反馈循环”，他们试图呈现出越来越多的“完美”表演，但当他们听到录音结果时，却一次又一次地失望。于是，表演变得不那么个性化，失去了自发性，开始更加标准化。听录音的体验也带来了另一种反馈，例如，小提琴颤音最初是一种留声机效果，但很快就被表演者模仿。

接下来的一次声音革命是1933年立体声留声机的发明。它是通过在凹槽的两侧分别录制互成90° 以及与铅直面成45° 的两个信道来实现的。立体声唱片的引入，意味着原则上可以再现整个原始表演的三维声场。这就产生了扬声器理想放置位置的问题，这一问题一直吸引着音乐爱好者。同时，也引出了“保真度”(fidelity)的概念，因为人们现在追求的效果是再现一个原始的表演。20世纪70年代之前，许多流行音乐作品从未像现在这样演奏过，当时音乐的大部分内容都是在乐队离开后添加的，这种情况随着后期处理和混合技术的进步而有所改善。然而，对古典音乐来说，保真的录音仍然是关键。尽管数以百万计的业余爱好者和专业听众、录音机、表演者和演奏者对“保真度”有着数十年的兴趣，但我们仍然无法对“保真度”进行量化。

存储的声音

当扩音技术和扬声器的设计得到完善之后，人们主要担心的是记录的脆弱性。一种关于如何处理、收藏、清洁，以及经过训练后正确播放录音的文化因此形成了。自动换片器的出现让一部分人很高兴，不过另一些人则将其视为最令人扫兴的发明，还有人认为它是新兴的

唱片破坏机。之所以会这样，一部分原因是自动换片器具有神秘感，另一部分原因是传统黑胶唱片，尤其是后来被称为专辑的**密纹唱片**(LP)质量极高，这让它们被尊崇的程度空前绝后，甚至至今仍有销售。

1964年，模拟盒式磁带作为一种强大又袖珍的唱片替代品被引入市场，并开始流行起来。尽管磁带很容易买到，但许多人宁愿买一张黑胶唱片，然后(非法地!)把它烧制到磁带上听，同时保持录音的原始精度。无线电广播也可以用这种方法录制，组合音响由无线电、盒式录音机和唱机组成，因为它录制盒式磁带时最方便，所以变得很受欢迎。

但是磁带有两大缺点，选择音乐时转带子(倒退或者前进)耗费时间，而且不可避免地发出高频嘶嘶声。杜比系统的许多变体在一定程度上减弱了嘶嘶声。这些系统都是在录制一个音轨时将高频信号增强，然后在回放时抑制整个音轨来生效。这是一种叫作压展(压缩和扩展的合成词)的技术。

立体声使人们对研究新的声音设备充满兴趣，并取得了很多新成果，磁带的制作也进一步发展。混音带就是磁带制作上的一个新成果，另一个新成果则是索尼随身听，它让音乐爱好者无论身在何处都能听音乐，不过，这有时会惹恼坐在旁边的人。

我们需要做的是完全抛弃模拟录音，也就是说，将声音以连续变化的形式存储(无论是物理的，如黑胶唱片，还是磁性的，如盒式磁带)。在数字系统中，信号被编码为一串 数字，以这种形式存储、传输或复制不会降低信号，也不会产生背景噪声。

乍一看，为了捕捉复杂声波的复杂性(如图9所示)，我们似乎必须测量和编码该声波的许多点的振幅。事实上，对声音的采样频率是希

望保存的最高频率的两倍就可以了。因此，要编码一个包含最高频率为8千赫的信号，我们就必须以16千赫进行采样(这就是众所周知的**奈奎斯特定理**)。如果以较低的速率采样，编码的数据就会失真，这就是所谓的混叠效应。

1982年光盘(CD)的推出导致了从模拟到数字的大规模转变。在CD上，数字编码的信号存储成光盘闪亮金属涂层上(塑料基片中)的深色凹坑，通过激光扫描即可读取，因为光盘上光滑的区域可以反射激光，但凹坑无法反射激光。CD播放机将反射翻译为1，非反射翻译为0，1和0的字符串将音频信息编码成二进制数字序列。

当然，如今购买、存储和播放音乐都是极为普遍的行为，甚至无须使用物理媒介。音频文件可以简单地下载到电脑上，通过各种设备播放。计算机通常是MP3播放器的一部分(MP3原意是“动态影像专家组音频层III”)。

MP3音频文件的神奇之处在于它们携带的文件非常小，大约只有CD文件的十分之一，这意味着一分钟的MP3音乐可以压缩到兆字节。能够压缩到如此令人吃惊的程度，是依靠一种叫作**霍夫曼编码**的技术实现的。在这种技术中，出现概率高的符号用较短的方式进行编码。另一方面，该技术对那些一丢失就引起人们极大关注的频段(主要是语音频率)会进行更加充分的编码，而对那些不受人们关注的频段则轻描淡写。

由于MP3播放器在决定歌曲压缩过程中需要去掉哪些信息时，同时考虑了音乐和听者，斯特恩得出结论：

> MP3包含了对交流、听或说、“心灵的耳朵”如何工作以及对音乐创作的实际与哲学层面的理解。每一个MP3的编码是由各种声音组成的一个世界，其中还有声音实践的整个历史。MP3编码器通过计算不断变化的录音内容与声音间的间隙，并据此构建声音间的时间对应关系来形成文件。MP3编码器工作得很好，因为它将想象中的倾听者定义为“非理想条件下的

不完美的聆听者”，猜测这些听众不会一直将注意力集中于他们正在聆听的声音，而编码器的猜测经常是对的。

高效的音乐编码利用了这样一个事实，即经过数十亿年的发展，我们的听觉系统已经进化到会对与我们最相关的声音做出反应。再加上声音本身的局限性，我们对声音世界的直接接触被限制在一个频率范围内，而这个频率范围只是实际存在的声音的一小部分。那些听不见的领域将是第6章的主题。

06 超声波和次声波 Ultrasound and Infrasound

蝙蝠的声音

许多教科书会告诉你，你的听力范围在20赫兹到20千赫之间，但你千万不能相信。如果你现在20多岁，那很可能已经对17千赫以上的声音充耳不闻了。随着年龄的增长，人类听力的高频极限下降明显且可预见。2009年起，愤怒的商店店主开始使用一种被称作“蚊声器”(Mosquito)的高频声波发生器驱逐未成年人，因为只有青少年(以及儿童和婴儿)还有能力听到这种设备产生的恼人音调。我们这些对17千赫的声音都充耳不闻的人则不会因为这些音调而感到不适，这表明能够听得到这么高的音调对演化并没有什么好处，否则我们早已经演化出更强健的听觉系统了。

尽管如此，我们还是错失了一种演化的良机，并不是因为我们错过了超声波特别丰富的声景，而是因为我们不能利用声音的一些便利的物理特性。这些特性在我们能听到的频率范围内是可以忽略的。然而，蝙蝠比我们要灵敏得多。

蝙蝠使用超声波的复杂程度令人震惊。在漆黑的夜晚，一只拥有80厘米翼展的埃及果蝠可以轻易地在两根垂直杆之间飞行而完全不触碰到杆，即便两根杆之间的距离只有53厘米。为了完成这样的壮举并展开捕猎行动，蝙蝠使用回声定位技术产生超声波并对它们听到回声的延迟进行计算，以获得附近物体的距离。

失明的人会根据回声时延来判断自己到墙壁的距离，但他们最终得到的并不是世界的视觉景象。考虑到可听声的波长，他们不可能得

到关于环境的全部景象。想象一个码头矗立在平静的湖面上，微风吹过，形成了大约1厘米波长的波纹。这些波纹一圈圈地从码头的柱子上反弹回来(见图21)。后来，大风来了，产生了更大的波浪，这些波浪从柱子上滚过，仿佛柱子不在那里一样。同样的原理也适用于声波，声波只受比波长更大的障碍物的影响，它也只能探测到这样的障碍物。

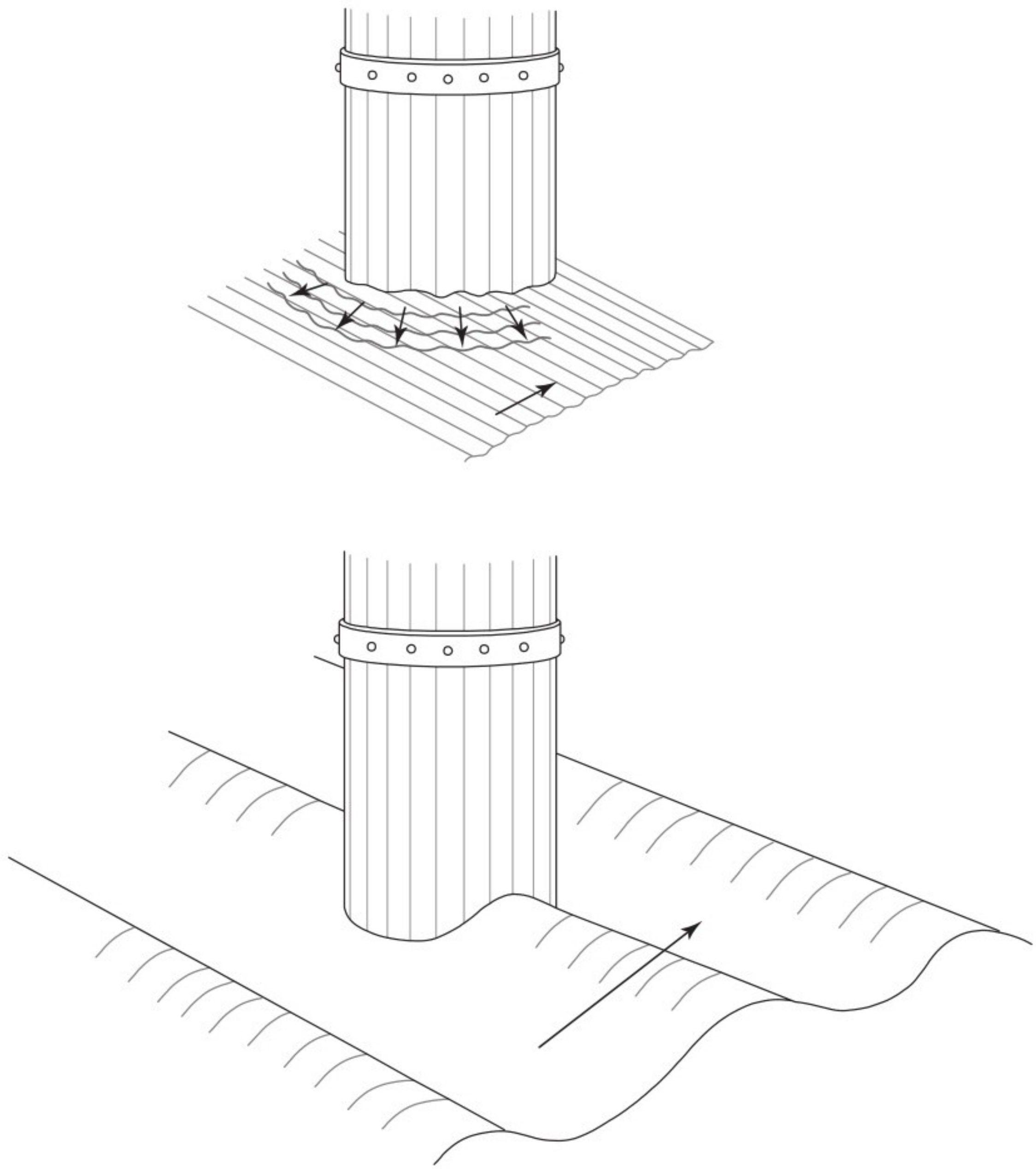


图21　涟漪与波

因此，如果一只蝙蝠试图靠回声定位一只蛾子，发出了一个人类可以听到的音调，比如中央C。中央C的波长是132厘米，只有当蛾子的大小超过1米时，蝙蝠才能够成功测到。要检测一只1厘米的蛾子，至少需要33千赫的声波。事实上，蝙蝠通常会发出80千赫的声音(较小的种类能发出更高的频率，范围在40～120千赫)。如此高的频率使得它们不仅可以处理孤立物体的位置，还可以处理树枝、树叶和昆虫的复杂三维分布。其中许多物体相互之间都在相对运动，而且都在相对于蝙蝠运动。

连续不断地发出超声波，对蝙蝠而言没有任何好处，其结果会是蒙太奇一样令人困惑的声波重叠和干扰性回声，无法被解码。蝙蝠所需要的是脉冲声，且声音必须是短暂的。如果两个飞蛾相距1厘米，想要产生独立的回响，声音脉冲必须在到达第二个障碍物之前就已经越过前一个了。因此声波的波长必须不到1厘米，而且持续时间不能超过30微秒。

蝙蝠或机器不可能产生单一波长的短时间持续脉冲，除非该波长比脉冲本身短得多。相反，蝙蝠发出非常尖锐的咔嗒声。根据傅里叶分析，我们知道这样的咔嗒声相当于不同频率的波的混合(见图12)，而且咔嗒声越剧烈(也就是说，它从安静状态变化到最大值的速度越快)，产生的频率范围就越大，这样就可以更准确地判断距离。

那对蝙蝠来说，最理想的声音是持续时间极短的一个咔嗒声吗？并不完全是，短时间的咔嗒声意味着声音的能量很低，那它就不能传播很远且不消退。人类在设计雷达和声呐系统时也面临着这个问题。1960年发明的频率扫描技术解决了这个问题，它利用了一个相对较长的频率逐渐增加的脉冲，并在相当长的时间内可持续产生大量的能

量，因此频率的范围也会比较大。而频率的变化意味着位于不同距离的物体的回声可以通过其频率的差异来区分。

当蝙蝠使用这种扫频技术时，脉冲持续2～3毫秒，在此期间频率下降(而不是像人类扫频器那样上升)约一个八度。蝙蝠产生这种脉冲越多，它收到的信息就越多，所以它通过修改咔嗒声的频率来应对不同的挑战。从寻觅猎物时发出每秒约10个脉冲到面临复杂环境或接近猎物时提高至以每半秒为周期发出100个脉冲。

如第2章所述，如果造成回声的物体在(相对于声源)运动，声音的频率将会因多普勒效应而改变。人类和蝙蝠的声呐系统利用这个效应来确定这些物体的速度，然而不同之处在于，我们的系统测量频率的变化，而蝙蝠修改它们的输出脉冲的频率，直到它们听到的回声不再变化，就仿佛目标是静止的。例如，如果一只蝙蝠正接近一个物体，则从物体返回的回声的频率不断上升，蝙蝠会因此降低它输出的频率，使回声的频率降低到与蝙蝠原来的频率相匹配的程度。

一种不同的声音

超声波和可听声的另一个物理区别是，超声波很容易形成声束，这对蝙蝠来说也是相当有利的。一个80千赫的音穿过1厘米的开口，将形成一个锥形声束，在1米距离之内扩散到约90厘米宽(见框9)。对蝙蝠这个以鼻孔发射超声波的物种来说，两个声源之间的干涉现象意味着声束之间的间隔进一步缩小。这不仅集中了声波的能量，使蝙蝠可以探测到更大的距离，还减少了分散注意力的侧面回声的数量。2015年一项关于蝙蝠接近饮水池的研究表明，蝙蝠接近水的时候嘴巴会张大，就很有可能是为了减少侧面回声的数量。

框 9

声音通过开口时的衍射：

衍射角度满足：$\sin(\theta/2)=\lambda/d$

其中，开口直径为d，波长为λ。

毫无疑问，自然演化为蝙蝠提供了用于生成和探测更高频率声音的生理装备，但空气对这些声音的吸收却是一个无法逾越的障碍。实验表明，在蝙蝠最喜欢的环境下(25℃，50%的相对湿度)，100千赫的声音每米有3分贝被空气吸收。也就是说，在1米的范围内，这种声音的声强会下降到原来的一半左右。相反，30千赫的声音只以每米0.7分贝的微弱速度被吸收，这意味着它们的声强下降了约15%。(声音的吸收随着频率的增加而增加，这也是雷声在近处听起来如爆竹般噼啪响，而在远处听起来却轰隆隆的主要原因。)

为什么会这样呢？因为空气是由分子组成的，所有的分子都在以一定的速度随机移动，并经常相互碰撞。在炎热的天气里，这种分子运动的速度会增大。事实上，温度只是对一大群分子运动速度的一种度量(见框10)。声波是一系列在空气中交替移动的高压和低压，所以在声波到达的任何特定位置，空气分子都会先聚集在一起，再分散开来，然后再次聚集在一起。当分子聚集时，它们会减速。就像一个人逆着人群的方向匆匆穿过拥挤的车站，人数越多，速度越慢。分子越多，速度同样也会越慢。当这种情况发生时，分子的能量就会发生变化。虽然分子移动得更慢了(换句话说，它们的动能下降了)，但旋转变快、拉伸变多，因此这些分子的内能反而上升了。

框 10

气体的绝对温度（以开尔文计）与动能$\frac{1}{2}mv^2$成正比，其中分子质量为m，分子速度为v。

这一情况可以类比为由坚固的钢弹簧连接的一对板球，也就是一个相当精确的双原子(两个原子)分子模型，就像构成了我们大气99%成分的氮(N2)和氧(O2)的分子。在通过声波的压缩部分之后，分子再次分离，转速变慢，拉伸程度下降，并以更快的速度运动。按下钢琴上的中央C，声波产生的能量以每秒262次的频率在动能和内能之间来回转换。

但如果频率足够高，能量转换的时间就会变得非常短，以至于一个分子无法在反转之前足够快地完成内能到动能的转换。结果就是，声速下降，声波迅速消失。这些变化开始时，声波实际的频率取决于它所处的介质，且固体和液体中的声波频率远高于气体中的声波。介质的其他特性，特别是黏度，也与这种效果有关。

这种现象对我们非常有用。当超声波衰减时，它的能量通过介质传播，并使其升温。这种热效应有许多应用，包括通过体内组织的升温以改善血液流动，或治疗受损的肌肉和关节。

医学超声检查

超声波还有很多其他的医疗用途，包括胎儿扫描(几乎所有发达国家的人都在出生前接受过扫描)和超声治疗(比如用25千赫的声波与水射流一起去除牙垢)。与其他许多医疗方法不同，超声波可以瞬间开启和关闭，且只需要低成本的技术，而且通常患者只需要最简单的准备工作。超声波发生器相对便携且只需少量辅助设备的特点意味着它们可以在医疗之外的领域使用，包括许多健身房中使用的透热疗法(深层加热)仪器，还有部署在战场上的伤口热凝系统——这个系统专门用以拯救可能因失血过多而丧生的士兵。

超声波已被用于治疗多种肿瘤，包括除此之外无其他手术方法的一些脑癌。这种技术被称为**高强度聚焦超声**(HIFU)或**高强度治疗超声**(HITU)。除了通过加热肿瘤组织(在米粒大小的区域里加热至约90℃)来破坏它们之外，超声诱导肿瘤中气泡的形成也是常见的疗法，超声诱导后化疗将更加有效。使用超声波来治疗这类需要精确定位的疾病的主要挑战是，超声波束的传播受限于人体组织的密度和弹性。因此，由人造组织模拟材料制成的人体部件模型被用来校准和编程设备。

超声波在医学上的一个更直接的应用是碎石术。在这种方法中，高功率脉冲只是在原位敲打肾结石，将它们粉碎成小到足以随尿排出体外的颗粒。

超声波扫描

超声波最著名的用途之一是扫描胎儿。由于传统扬声器的膜片移动速度不够快，无法产生合适的兆赫频率，因此使用了压电发射器。将凝胶涂在腹部，这样就没有空气层来反射或吸收声音。超声波在具有不同声阻抗的介质(如骨骼、肌肉、皮肤、羊水等)之间的界面反射，通过精确测量回声从这些界面返回所需的时间(每种介质中声速不

同)，就可以计算出它们的距离。通过移动超声波束，可以测出详细的三维图像，并将其转换为实时视频图像。

像大多数扬声器一样，压电换能器是由它产生的振荡电流往复产生声波。另一方面，当被声波击中时，它会产生电信号。因此，在胎儿扫描仪中，扫描仪的探头既是超声波的来源，也是超声波回声的检测器。

利用极高频声波束可以生成相当精确的图像。1兆赫(100万赫兹)的信号可以显示毫米尺度的图像细节，而检测结果的精确度取决于被检查组织的声学性能。并且，现在许多扫描仪已上升到15兆赫，眼睛和皮肤用的扫描仪甚至可以达到50兆赫。

但与声学显微镜产生的80亿赫兹信号相比，这简直是小巫见大巫，因为声学显微镜可以观察到0.03微米尺度的细节。但不幸的是，如此高频的声音还没传播到1毫米外，就会被几乎所有类型的介质吸收，唯一的例外是液态氦。注意，如果没有保存在温度低于5开尔文(-268℃)的环境中，液态氦就会沸腾。由于需要精密冷却系统，声学显微镜的价格会比较昂贵。即便如此，我们还是要使用声学显微镜，因为它们可以探测到样品的表面以下，特别是一些难以从视觉与周围环境区分，但对声音有很强反射的材料。

在医学领域之外，超声波最常见的诊断应用之一是检测缺陷和裂纹，例如在铁路修建方面。为了定位这些缺陷，人们沿着被测试的物体发送一系列的音爆。在脉冲回波模式下，发射器和接收器被放在一起，如果存在缺陷，脉冲被反射，它们的到达时间能够表明缺陷的位置。在传输模式下，探测器和发射器是分开的，传输过程中脉冲的任何变化都表明测试对象中存在不均匀性。这种方法也可用于测定固体中的机械应力。由于材料的弹性模量在受到应力时发生变化，它们的声速也会发生局部变化。

超声波的力量

虽然大功率超声的热效应有许多应用，但它也可以通过对介质的机械作用来传递能量。枪虾就是一个在自然界中比较罕见的例子，利用超声波(连同可听频率)产生的冲击和压力来捕猎。当这种动物猛夹它的虾螯时，会发出响亮而突然的咔嗒声，其频率高达200千赫，足以杀死或击晕猎物和潜在的捕食者。

无论是虾还是人类，超声波的机械能通常是通过**气蚀**传递的。气蚀是指微小气泡的形成和剧烈坍塌。基本上所有的液体都含有这种气泡，要么由它们自己的蒸汽构成，要么由空气构成。当液体的压力下降时，这些气泡就会膨胀(这就是打开一个装有汽水的加压容器时产生泡沫的原因)。因为声波是由一系列的高压和低压组成的，它会使气泡迅速膨胀和收缩，在高功率和高频率的情况下，气泡的振动非常剧烈以至于产生破裂和内爆，同时以热能的形式释放出它们的振动能量。其产生的温度甚至可能会超过太阳表面的温度，并能使液体发光(一种被称为声致发光的现象，枪虾也能引发这种现象)。

因为能量只会在很小范围内出现，所以液体不会全都变得特别热。但是，高浓缩能量的爆发可以用来引发化学变化(声化学)，或者清洁和消毒需要在水下清洗的物体，比如某些医疗器械。基于气蚀的清洗在20～50千赫范围内最为有效。在更高的频率下，超声波还会搅动液体，在100千赫到1兆赫的范围内，液体搅动在清洗效果中占主导地位。在实际生活中，超声波清洗浴缸就同时应用了这两种原理。

高功率超声(无气蚀化)常常被用于电路板印刷中的无焊剂焊接，其中电烙铁的电热头以超声波频率振动。在更高的功率下，非常精细的线可以通过由超声波振动引起的摩擦从内部加热而焊接在一起。这样做的一个好处是，由于加热效果仅限于非超声(充声)电线，超声波

不会加热附近的元件。还有一些其他依赖超声波焊接的材料，其频率的选择取决于要焊接的零件的大小，范围从小零件的60千赫兹到大零件的10千赫。

高功率超声波甚至可以用来举起小物体。虽然它们所产生的力是微弱的，但在微重力环境中，例如在空间站内部，它们可以在组装过程中使精密仪器的位置保持稳定，或者防止高度活跃的化学物质接触到任何设备。

外部极限——声子

在极高的频率下，声音的表现形式与我们所熟悉的完全不同。就像电磁辐射，最高频率的电磁波的表现形式不像是波而更像粒子(因此盖格计数器可以通过一个个的咔嗒声记录下伽马射线的存在)。表现形式更像粒子的那些最高频的声音被称为声子。人类是间接发现声子的存在的。在19世纪末，人们已经知道要使一种物质的温度升高1℃，需要特定的热量。水的比热容比油的比热容高，这就意味着烧一壶水比烧一壶油要花更长的时间。

气体和固体也有特定的热量，但有一种异常现象令人不解。提高固体的温度所需的热能大约是提高相同数量的气态物质温度的两倍。这意味着固体(和某些液体)一定有某种储存热量的方法，而这种方法是气体无法获得的。这个原理就是振动。固体中的分子可以像钟摆一样围绕其平衡位置振动。但是，与钟摆不同的是，振荡分子不能逐渐变慢。量子力学定律要求它必须从快速振动跃迁到缓慢振动。当它这样跃迁的时候就会把一个声子的振动能量传递给另一个分子。固体传导热和传导电的方式可以通过声子的行为来解释。

次声波

随着声音不断接近人类听觉频率的下限，声音刚好能被听到的声强就会升高。正好能被听到的20赫兹纯音比正好能被听到的4千赫声音的声强(在其他条件相同的情况下)高出近3亿倍。这种强大的低频声音在空气中很少见，但在建筑工地、地铁站、高速公路附近以及地震活跃地区，我们很容易就能感觉到这种由固体传播的声音。

另一方面，通过空气传导的低功率次声波一直环绕着我们。它们甚至可以在我们走路的时候产生。由于我们头部的上下运动，耳朵的循环气压变化在我们周围形成了一个约1赫兹的次声波。海浪能够产生0.2赫兹左右的次声波。所有声音中频率最低的自然声音都来自高空和地下深处，极光和火山产生的次声波约为0.01赫兹。

次声波的主要特征之一是它的传播距离比可听范围内的声音更远，并能穿过海洋、地面或空气。在空气中，几千千米外的次声波很容易被探测到(例如，由火山爆发产生的次声波)，尽管通常不是靠麦克风，就像我们通常是靠感觉，而不是靠听觉来发现次声波。我们使用专门的气压计来测量次声。次声也可以通过它引起的温度变化来检测。

有证据表明，次声对人类的影响与其他类型的声音明显不同，其中就包括增强情绪反应。现代古典音乐会加入次声，同一支乐曲在一场没有次声伴奏的音乐会中会有更多的人更讨厌或喜欢。另一些实验也证明，驾驶员暴露在次声环境下，很快就会感到非常疲劳。甚至有人认为，次声是使人们产生幻觉的原因之一，部分原因是直接的情感影响，部分原因是眼球的振动会引起视觉障碍。

次声已被证明是探测火流星的可靠手段。火流星是一种在飞行中爆炸的流星体。由于它们在大气中超音速飞行，下落时会产生富含次声波的音爆。更多的次声波是在火流星爆炸时产生的。所有这些空中的次声到达地面时会产生固体中的次声，当火流星碎片撞击地球时会

产生表面波。将所有这些声音的频率、时间和振幅结合，我们就可以对火流星的路径、运动和能量进行详细的分析。

次声在地面传播时对我们的影响更大，而超声在水下传播时才真正地如鱼得水。我们将在第7章中介绍这些声音传播的媒介。

07 水下和地下的声音 Sound Underwater and Underground

第二种视力

如果我们的世界一直笼罩在雾中，可能我们就会极少使用视觉，而更多依靠听觉。假设这种雾是无线电波无法穿透的，那我们的通信系统无疑也会在声学上发生改变。海洋就是这样一个世界，水中有许多悬浮粒子，能见度很低，且无线电波在其中很快就会消失。但水里充满了声音，鱼类、海洋哺乳动物、人类潜水员、潜水艇和水下机器人，它们都用声音来进行微妙而复杂的感知、交流，甚至有时还会利用声音发动攻击。

这对不到100年前的科学家来说是闻所未闻的。实际上，列奥纳多·达·芬奇(Leonardo da Vinci)在1490年左右指出：“如果把你的船停下来，将长管的一端放在水中，另一端放在耳边，你就能听到离你很远的船的声音。”但当时没有人注意到他的这一论述。

在第一次世界大战期间，人们才开始尝试倾听船在水下的声音。而直到第二次世界大战，水下声景才变得丰富起来。这在一定程度上是由两个奇怪的事件引发的。1942年，部署在切萨皮克湾(Chesapeake Bay)用于探测德国潜艇的声学浮标全部同时发出警报。一队驱逐舰“满怀热情”地向这片海域发动了深水炸弹袭击，但没有发现任何泄露潜艇踪影的油迹，受到攻击的只是成千上万条鱼。同年晚些时候，在美国西海岸，所有的感音水雷(这是为了保护港口不受任何带螺旋桨的东西的袭击，以防日本入侵而安装的)同时爆炸，结果造成大量海洋生物死亡。

如果美国军方对一种叫黄花鱼的鱼类有更多的了解，就可以避免许多类似的悲剧。尽管外表不起眼，但这种棕色的小动物拥有非常响亮的声音，就有点儿像放大后的啄木鸟声音。就像鸟儿一样，每天清晨这些鱼会聚集在一起发出声音。

鱼类的嘈杂几乎完全出乎生物学家的意料。一直以来，有一种深入人心的误解，即海底就像人们传说的那样寂静无声。事实上，虽然海洋经常被开发用于沿海和国际旅行，并作为人类食物的来源，但除了道听途说的有关海怪的奇闻和沉没城市的传奇故事之外，人们对海底的东西基本上一无所知，甚至很少有人去推测。黑暗的水下世界对人类是完全陌生的，似乎没有人类可以涉足的地方。甚至水手也把深海视为禁区，他们中的大多数人甚至没有学会游泳。虽然艺术家、作家和心理学家常常把海洋看作原始自然的狂野领域，但他们的注意力却仅仅局限于它粗糙的表面。在这个表面之下是一个不活跃的区域，不见天日，暗淡无光。

20世纪40年代初，人类开始使用水听器(水下麦克风)聆听并揭示了“寂静的深渊”到底有多吵，探测到人类听力频率范围以外的各种各样的声音。声音的来源多种多样，而且其中大多数是无法识别的。在一些地方，最大的声源竟然是枪虾群。事实上，上层海洋区域从来都不是平静的，它的背景音由波浪、雨水、闪电以及海洋生物产生。在浅水区，旋转纷飞的沉积物是一种额外的、无处不在的声音来源。

听到水下的声音

表面上看，海洋的静默有两个原因，一个是物理上的，另一个是生物学方面的。物理上的原因是空气和水之间的声阻抗不匹配，海洋表面就像一个声波反射镜，几乎能将海面下的所有声音反射回去，所以陆地上的居民只能听到海浪的拍打声。

将你的头浸入水中，生物学上的原因就显现出来了。在水下，耳膜的一边是水，另一边是空气，因此声阻抗的不匹配再次阻止了大多数声音进入耳朵。如果我们没有鼓膜(也没有充满空气的中耳)，我们就可能在水下听得很清楚。

水下动物不需要人类这种复杂的耳朵。由于周围水的密度和动物身体的密度相近，声音进入和通过它们的身体都很容易，水下动物不需要耳郭来收集声音，也不需要鼓膜或类似鼓膜的圆窗结构来将它从一种介质转移到另一种介质。鱼确实有被称为耳石的耳骨。其成分是碳酸钙，碳酸钙的高密度提供了足够的声阻抗差，让声波引发振动。这种运动传播到生长在毛细胞上的立体纤毛上，向大脑发送神经信号。其他被称为神经丘的毛细胞分布在鱼的皮肤上。

还有两个结构也可以增强某些鱼类的听力。第一个是鱼鳔，它是一个充满空气的囊，功能类似潜水艇的压载水舱，根据需要改变所受浮力从而使鱼可以不费力气地下沉或上升。鱼鳔作为一个相当灵敏的听觉器官，很容易感受到声波的振动，它能听到3千赫左右的声音。然而，作为一个单一的、对称的器官，它有一个很大的缺点，即鱼鳔不能提供声音的方向信息。

这个缺陷由第二种结构弥补，即侧线，一个沿鱼身两侧分布的充满液体的管道，作为在低频(160～200赫兹)环境下的一个方向敏感的声音探测器。人耳静纤毛在声波引起基底膜振动时随之摆动，不同于此，侧线静纤毛直接被传入的声波推拉，这意味着声音的方向可以被直接感觉到(也就是说鱼能感觉到水分子的运动，而不像我们一样通过压力变化来感知)。这使得人类很难悄悄地接近鱼而不被发现。

各种技术和传感器

最早在水下使用声音且能被称为“技术”的是贝尔系统。在这个系统中，放置在港口附近的水下钟可以被装有原始水听器的船只探测到。这种原始的水听器采用带有防水外壳的碳质麦克风。一个看守人在船上用立体声监听这套系统，这样即使在能见度很低的情况下也能把船引导到港口。在1875—1930年，这个系统被安装到许多船只上，包括“泰坦尼克号”和“卢西塔尼亚号”。到1923年，英国海岸还有30个水下钟。但从大约1910年开始，该系统逐渐被回声测深技术所取代。回声测深技术是在水下发出声音并对回声进行计时，当水下声速已知时就可以计算出距离。这也是脉冲回声技术的另一个例子。

从探测回声开始，现代声呐(声音导航和测距)系统逐步发展起来。在主动声呐系统中，持续时间短的声音脉冲被船只投射出去，在它所遇到的物体上反射。除了确定距离，我们还能靠收到的脉冲频率的变化来计算声源的相对运动速度(借助多普勒效应)。

被动声呐系统只是简单地监听水下声音，尤其是船舶发出的声音。自动声音识别技术可以通过发动机的声音甚至是电子系统的嗡嗡声来识别不同类型的船舶。事实上，每艘船舶都有其显著特征或声学指纹。这种技术在冷战时期对于识别及跟踪敌舰和潜艇具有非常重要的意义。主动和被动声呐系统有时被部署在浮标(声呐浮标)上，浮标装备通过搭载的无线电系统来报告它们探测到的东西。

水听器是水声工作的关键仪器。现在使用的几乎都是压电材料，通常是基于一种叫作PZT(铅和锆钛)的合成陶瓷。与麦克风不同，水听器有时必须具有非常大的规模才能在低频率下具有方向性。因此一部分潜艇的侧面几乎完全被水听器覆盖。这些所谓的“侧翼阵”通常由聚偏氟乙烯(PVDF)制成。

要在水下使用扬声器功能，就需要水声发射器。与扬声器相比，水声发射器有一个局限性，即当声波由水面向水下移动时，声波会变

得稀薄，如果压力稀薄到低于周围水的压力，就会产生空化现象。气泡分散并吸收声音，使水声发射器静音。深度越大，水压越高，水声发射器在空化发生前需要产生的声功率就越大。

在高功率的应用中，如石油和天然气勘探类的地球物理测量中，气枪被用来产生水下声音。在这些情况下，一个小的空腔充满压缩空气，由于继电器突然释放压力，空腔迅速形成一个大气泡，伴随而来的是巨大的爆裂声。脉冲频率在20～200赫兹，振幅也非常高，这可能是除大型爆炸外，海洋中最“响”的人造声源。声音穿过海床，从下面岩层之间的界面反射回来。利用非常长(长达10千米)的水听器阵列在水面附近对反射的声音进行成像，最终观测结果被计算机处理后可以形成一个三维地图。

虽然水声发射器在原理上只是反过来的水听器，但它们的物理设计往往是不同的。最广泛使用的类型就有蘑菇[1]形压电换能器(Tonpilz transducer)。几个PZT压电陶瓷片在电极之间夹成一叠，末端是锥形或圆柱形的质量块，也就是水声反射器的工作端。Tonpilz压电换能器可以产生2～50千赫范围的频率。

对于包括信号和距离传感在内的许多水下应用，常常需要借助具有方向性的声音。就像在空气中一样，如果一个声源产生的声波波长小于换能器表面的宽度，那么它就会自然地具有方向性。但是因为声速在水中比在空气中要大5倍，那么特定频率的波长也比它在空气中对应的波长大5倍，所以更难判断其方向。

一种制作定向声源的经典方法是使用参数阵列。如果两个声源产生的波只有频率相差一点，那么就会产生差频，以及频率为声源频率之和的波。差分波的波长根据需要可以为任意长，但它保持了其生成波的方向性。

参数阵列利用了声速依赖于密度这一事实。在高声能下，压缩体中的压力变得非常大，密度显著增加使得声波会短暂加速。密度变得稀疏时，情况则刚好相反。这些速度变化的影响会扭曲波通常的正弦形式。

这在高功率超声波中也很常见。正如傅里叶指出的那样，非正弦波等价于分量正弦的和。考虑参数阵列的情况，这些分量包括原始波，以及加和波与差分波，其中差分波是我们感兴趣的一种波。参数阵列也可以在空气中使用，可以使音频听起来更具方向性。

电磁波在水面以上可以完成的大部分任务，在水面下都可以通过声波完成，但声音有一个不可避免的缺点，即它在水中的速度远低于空气中电磁辐射的速度，这意味着扫描需要更长的时间。同样，当波被用来发送数据时，数据传输的速率与波的频率成正比——而音频声波的频率大约是无线电波的千分之一。由于这个原因，超声波成为一种替代方案，因为它的频率可以与那些无线电波的频率相匹敌。超声波的另一个优点是更容易产生定向波，且只向你想要的方向发送信号。然而缺点是吸收比例随声音频率增加而增大，所以频率范围是有限的。

传遍全球的声音

声音在水下传播的距离是惊人的。据称，在发动机驱动的船只大量出现之前，南极的鲸类能听到它们远在北极的同类的声音。声音能够传播到如此大的范围，部分原因是声波在水中的吸收要比在空气中少得多。在1千赫时，空气(湿度为30%)的吸收率约为5分贝/千米，而海水的吸收率仅为0.06 分贝/千米。此外，声波在水下受到的限制比其在空气中要大得多。在空中发出的噪声向四面八方传播，但在海洋中，海床和海面限制了垂直传播。

水下声速变化范围也大得多。因为温度、压力和盐度的影响，密度会产生巨大变化，从而影响声速(见框11)。随温度变化而引起声速极速变化的海水层被称为温跃层，在大多数海洋中温跃层的形式都是类似的。天气晴朗的时候，海洋最上层区域的特征是温度随深度增加而迅速下降，因此声速也随深度增加而下降。正因为晴朗的天气在夏天更常见，因此形成了众所周知的季节性温跃层。在此之下是主温跃层，温度和声速随深度增加而继续下降，与季节无关。主温跃层底部(其深度随纬度变化较大)温度稳定在4℃左右，再深处温度也基本没有变化。在这个深部等温层中，压力成为决定声速的主要因素，声速随深度的增加而增加，如图22所示。

框11　水中的声速

最精确的方程是勒罗伊－罗宾逊公式：

$$c=1\,402.5+5T-0.054\,4T^2+0.000\,21T^3+1.33S-0.012\,3ST+0.000\,087ST^2+0.015\,6Z+0.000\,000\,255Z^2-0.000\,000\,000\,007\,3Z^3+0.000\,001\,2(\Phi-45)-0.000\,000\,000\,000\,95TZ^3+0.000\,000\,3T^2Z+0.000\,014\,3SZ;$$

其中，C是声速，T是温度（℃），S为盐度（百分比），Z为深度（米），Φ为纬度。

所以在海洋中的某个地方总有一个声速较低的层，夹在声速较高的区域之间。由于折射作用，来自上面和下面的声波都被转移到这个

声速最小的区域，并被困在那里。这就是深海声道(deep sound channel)，它像一个薄薄的球形外壳，延伸在世界各地的海洋之中。

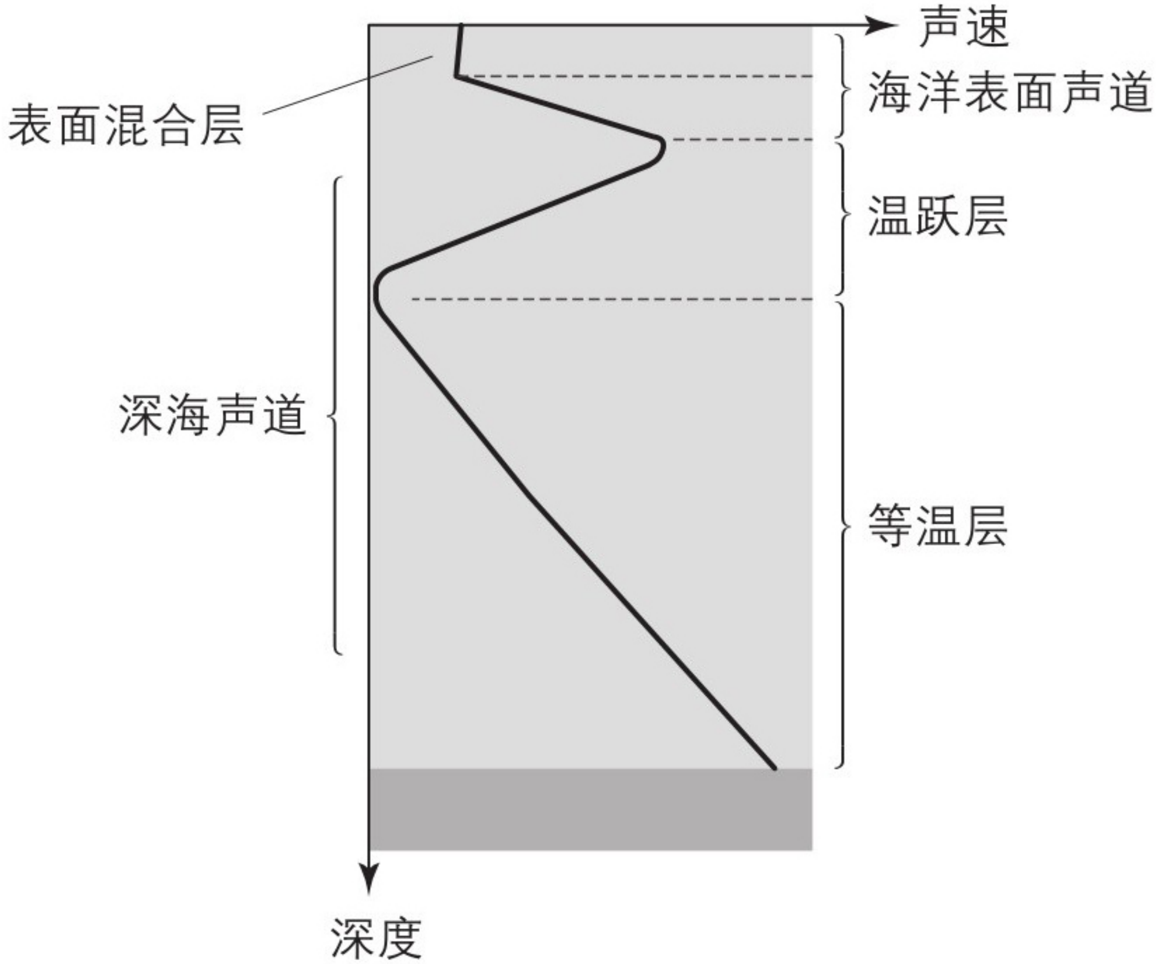


图22　水下声速图

由于声波在深海声道只能水平移动，其声强下降程度只与声音传播距离成正比，而不是距离的平方，这是因为它们只会在同一温度的空气或水中传播(换句话说，它们以环状而非球面分散)。深海声道的吸声率非常低(对频率有很强的依赖性，但对于4千赫波，吸声率约为0.2分贝/千米)，所以深海声道的声波可以很容易地环绕地球传播。

深海声道被用来建立**声音定位和测距**(SOFAR)系统。该系统于1960年由澳大利亚-百慕大声波传输实验发起，在澳大利亚靠近印度洋一侧海岸上的赫德岛附近引发了爆炸。爆炸的声音在距离2万千米之遥的百慕大被探测到。SOFAR系统的研究人员还发现了一种新的、意想不到的声音，后来被确定为长须鲸的叫声。长须鲸在很久以前就发现了深海声道的存在和特性，并定期造访深海声道，借此向它们遥远的同伴发出信号。

SOFAR系统为**海洋气候声学测温**(ATOC)系统开辟了前路。ATOC系统通过测量大范围的平均声速来计算全球海洋温度，从而帮助量化气候变化。

天气的变化导致海洋环境变化，从而导致海洋出现一系列暂时的声学异常状态，包括将大多数声音排除在外的阴影区，以及允许长距离传播的暂时的声音通道。由于在行进过程中速度的反复变化(这是由于温度和盐度的变化)，沿着暂时声音通道传播的声音就像那些在深海声道传播的声音一样，在传播过程中高度失真。在20世纪90年代，各种来历不明的声音非常怪异。其中最著名的可能是**海洋怪声**(Bloop)，它使人们产生了一系列丰富的想象。

然而，海洋怪声最有可能是来自远处的冰山扭曲崩裂的声音。

地球的声音

声波很容易在固体中传播(见框12)，其他类型的压力波也是如此。但并非所有的地震波都是声波。P波(初级波)是纵波，即一系列压缩波和稀疏波。其速度由地面的密度和弹性决定，因此是声波。然而，S波(次级波)是横波，因此不是声波。P波和S波都是震波，它们在

地球上传播，被地下的地层折射，为我们提供了有关地球结构的信息。地表地震波也多种多样，但没有一种是声波。

> **框 12**
>
> 在气体或液体中，声音的衰减可由斯托克斯定律给出：
>
> $\alpha = 2\eta f^2 / 3\rho v^3$
>
> 其中，α 为衰减幅度，η 为动态黏滞性，f 为频率，ρ 为密度，v 为声速。

许多大型动物能够发出并听到低频的声音，比如非洲象，原因很简单，它们的声带很大，移动相对缓慢，它们的叫声中甚至有一些是次声波。这对它们来说是相当有利的，因为次声传播远而衰减小(非常粗略地说，空气中10赫兹的信号传播距离比100赫兹的信号远100倍，比1000赫兹的信号远1万倍)。

声音在地下的衰减情况是极为多变的，但通常比在空气中要低得多。母象利用次声波来吸引雄象(空气中可传播3千米以上，而在地下可传播10千米以上)，并与它们的幼象联系。大象还能利用次声来探测500千米以外的雷暴(找到雷暴就能找到水源)。2004年，斯里兰卡的大象逃离海岸，可能是因为它们听到了即将到来的海啸产生的次声波。大象通过发出隆隆声或跺脚来产生信号，而大象跺脚时也可以通过一种叫作“帕西尼安小体”的振动感受器来探测地面的声音。

人类对地下声音的利用至少可以追溯到公元132年，当时中国制造了第一台地震探测器——地动仪。地动仪是一个有一圈金属青蛙(蟾蜍)的黄铜容器，每个青蛙嘴里都衔着一个金属球(铜球)。地震的发生和方位是由某个方向上的球体下落来表示的。

从远古时代起，地下的声音就一直是人们恐惧和惊奇的来源，在演化的进程中，这种感觉可能已经在我们的大脑中根深蒂固。从我们祖先的经验来看，这些声音伴随着雪崩、火山爆发、地震和其他有极强破坏性的自然灾害。它们既能被感觉到，又能被听到，这进一步增强了它们的影响力。感觉到坚实的地球在脚下移动和颤抖的确是一种令人不安的体验。难怪在许多文化中，地下的黑暗世界被认为是死者的住所。

然而，尽管人们对地下声音一直有负面印象，它却一直被用于战争中。在早期的战争中，地道中的敌人有时会因为他们在地下发出的声音被发觉。甚至有记录表明，人们可以通过将盾牌砸向地面时发出的声音来判断地下是否有地道。在第一次世界大战中，利用从战壕墙壁的不同位置探测到的信号，可以通过三角测量来估计敌军在自己战壕中的位置。在第二次世界大战中，波兰抵抗运动花了很长一段时间在地下监听和跟踪头顶上的德国士兵，

而德国士兵也在用同样的方法跟踪他们。

地下声音探测器被称为地音探听器(geophone)。直到20世纪晚期，地音探听器的工作原理还与动圈麦克风相同，磁铁因地面波而移动，从而产生绕其周围的感应电信号。如今MEMS成功取代了动圈麦克风(见第5章)，一块微小的硅被安装在一个精致的支架上，由于地下声音的振动而开始移动。反馈系统会停止这种运动，它所施加的力会提供关于声音的精确信息。但MEMS设备相对来说不太敏感，只能用于监

测活跃地震带。所有实际应用的检波器都是高度定向的，通常被用来响应来自地面正下方的声音。

勘探

勘探通过利用地下声音得到了长足发展。通常在地下引爆炸药的时候(在地下没有类似扩音器的装置)，它所产生的波会在不同材料层间的界面上反射，并被一组地音探听器探测到。

固体中应变产生的超声波被称为声发射，用对它进行的测量和定位来检测各种结构的裂缝。其应用范围包括检测飞行中的飞机和伦敦眼[2]。这就是所谓的结构健康监测。即使一个看不见的小裂缝刚刚开始形成，也可以在瞬间触发警报。声发射还被用于研究焊接过程中裂纹的形成，检测携带高压流体的管道中的损伤，并确定钢筋混凝土内部的腐蚀量。

地下核试验很难用其他方法识别，它产生的独特声音大多是次声波。**全面禁止核试验条约组织**(CTBTO)不断使用地音探听器(连同深海声道的水听器，探测来自海底岩石的水下声音)监测地球。这一地下和水下探测器网络使CTBTO得以监督国际禁试条约，一旦核试验的确发生并被监测到，人们就能计算出试验地点。

这一章集中讨论了我们听不见的声频区域，尽管许多科学家希望我们能听到这些声音。在第8章中，我们将注意力转向那些我们不得不听到，却又不想听到的声音。

[1] Tonpilz在德语中是“唱歌的蘑菇”。

[2] 伦敦眼是世界上首座、同时截至2005年最大的观景摩天轮，为伦敦的地标及著名旅游观光点之一。

08 纷扰之音 Sound out of Place

噪声的性质

噪声是什么？这是类似“时间是什么”的大问题。对于这种问题，有人可能会回答：“在被人问到之前，我还以为自己肯定知道。”更糟的是，“噪声”在两个不同的领域有完全不一样的“身份”。在科学领域，噪声是一种外界的声波或电磁能量；在通信领域，噪声是指任何不是信号的东西，换句话说，它不携带信息。信息可以是通过电话线传播的声音，也可以是超声波扫描中胎儿的身体结构。噪声的量越大，就越难将信号从中分离出来，因此就有了信噪比的概念。

一旦识别出噪声，消除它就相当容易了。发生在特定频率的电路中的噪声(例如电源发出的50赫兹的嗡嗡声)通常可以用滤波器处理，因此可以只保留有意义的频率。但坏消息是，电子学中一种非常常见的噪声形式是白噪声，它充满了所有我们感兴趣的频率范围。当它通过扬声器到达人耳时，听起来就像嘶嘶声。这样的噪声没办法轻易滤除，但人们可以滤除携带信号的频带以外的所有频带。虽然在信号波段仍然存在白噪声，但总体信噪比将有所提高。

然而，对大多数人来说，“噪声”是指任何听者不想听到的那些声音。一个号手可能会对自己演奏出的声音非常满意，当然不会把它当成噪声，但不管从技术上看他的演奏有多么优秀，邻居都很可能认为这是噪声。双方都是对的，一个人的音乐可能是另一个人的噪声。这就触及了“为什么噪声如此难以消除”这个问题的核心。幸好，几乎每种声音都有至少一个人喜欢它。但对有些人来说，周围环境必须安静，几乎任何声音都是一种诅咒。的确，对我们大多数人来说，听

起来是噪声的声音有三个共同的特征——猝不及防、响亮、无音调。为了找出这些共同特征的原因，我们需要思考为什么我们能听到声音。

祖传的遗产

我们有听觉，是因为我们的祖先有听觉；而我们祖先有听觉，是因为没有听觉的那些“人”被自然淘汰了。在我们祖先可以听到的诸多声音中，有几种与生存相关。例如：突然的咆哮声、弓弦“砰”的声音、脚下树枝的断裂声、头顶的雷声，这些都是危险的信号，而它们的共同之处在于突然性，这正是今天我们对突发性噪声有即时反应的原因。就像小鸟看到捕食者时会突然停止鸣叫一样，突然的寂静与突发的声音有同样的意味。因此，在时钟停止滴答，空调关闭，或者持续的雨突然停止时，就会无比寂静，而这种寂静很常见。事实上，我们周围任何意想不到的声音变化都可能是一种恼人的噪声。每个司机都知道，一种无法识别的新的引擎噪声会造成多大的愤怒和担忧，不管司机多努力地去提醒，无感的乘客根本无法从背景音中分辨出这种噪声。

但是，为什么我们会对那些似乎与我们的祖先面对的任何危险信号无关的声音感到烦恼呢？比如半夜屋外街上聊天的声音，或者在火车车厢内其他人播放音乐的声音。

答案在于声音的一项关键社会功能：展示力量，尤其是对空间宣称所有权。我们习惯于认为每个人都有一个如影随形的私人空间，不请自来的人是不受欢迎的。如果有人用他的声音侵入了那个空间，就像有人真正侵入空间一样令人讨厌，这两个例子的重点都是声音的目的性。如果坐在你旁边的人打电话给别人却净说些废话，而不是接到一个真正重要的电话，你是不是会觉得他更加惹人讨厌？

一个人是否会让别人觉得吵闹也取决于他和听者之间的关系。例如，历史学家肖恩·怀特(Shane White)和格雷厄姆·怀特(Graham White)在对美国奴隶制的研究中发现：黑人福音传教士的布道虽然被教会的会众认可，但对美国白人基督徒而言却完全是噪声，这是因为他们布道的声音和信息的传播范围远远超出了他们自己的圈子。另外，文化中的礼仪规范也很重要，在不同的国家，图书馆、就餐时间和葬礼上可接受的声音响度水平差别很大。

公共交通、公共空间或开放式办公室都可能成为令人生烦的环境，人们经常淹没在他人的嘈杂声中。个体的声音既不能被识别，也不能被定位，但听觉系统还是不断地试图同时完成识别和定位这两件事情，这使得人们既难以忽视这些噪声，也难以逃离这个环境。即使一个人能够有意识地忽略这些声音，其心理情绪也不会中断，而且会不断产生压力，甚至还会产生由这种压力引发的高血压。

通常，重新获得私人声音空间的唯一方法是用自己的声场来填充，比如MP3播放器或智能手机。英国萨塞克斯大学媒体与电影教授迈克尔·布尔(Michael Bull)说："iPod的泛滥可以被简单地理解为一种令人愉快的中毒形式，在这里，用户所处的'完全和解'的世界是一个不曾被入侵的梦想。就是说，iPod能够同时直接介入世界和人的情感之间。"

有些声音是不能被忽视的，即使它们并不存在。这些所谓的"耳虫"通常是一些简短的音乐片段，大多是陈词滥调、靡靡之音。它们非常顽固地在人们的脑海中不断重复。广告使用的那些短曲尤其善于钻空子。"耳虫"虽然相当普遍且看起来很简单，但实际上很难找到科学上的解释。神经学家奥利弗·萨克斯(Oliver Sacks)认为，人们记住音乐的那些不同寻常的方式可能发挥了重要作用。回忆一个场景或事件涉及重建它，这意味着回想起来的内容会因人而异。然而，在

回忆一段音乐时，一些更接近于直接复制原作的东西被保存了下来，用萨克斯的话说，这是一种“几乎没有防御能力的刻印在大脑中的音乐”，且我们对它的控制力相对较弱。也许正是因为音乐，以及由此而来的“雕刻”过程在人类心智的进化过程中出现得如此之晚，才使得这种记忆难以控制。

音乐对大脑的影响更为深远。对一些人来说，音乐可能会导致癫痫发作；但对另一些人来说，音乐有明显的镇定作用，甚至可以减轻疼痛和高血压。音乐在治疗一些心理问题上特别有效，自第二次世界大战以来它就一直被用来治疗心理疾病。音乐也被证明对缓解一些帕金森病患者的症状有很大作用。

震耳欲聋

实际上噪声引起的听力损失问题是很严重的，它造成的损害可能很多年后才会表现出来。此外，我们能够对危险的高温或强光立即做出反应，但对噪声的防御反应却要弱得多。

我们的眼睛具有一系列的保护适应能力，最明显的是眼睑和收缩的瞳孔(或扩张的虹膜)。为什么耳朵什么都没有呢？没有耳盖的原因是，虽然失聪可能会缩短动物的寿命，但这种寿命的缩短与被一只饥饿老虎成功伏击而立即毙命显然不能相比。我们的耳朵有一个和瞳孔相当的东西，虽然不是特别有效。听觉反射是由每只耳朵里的两块肌肉提供的，它们分别叫作镫骨肌和鼓膜张肌。当一个巨大的声音到达时，镫骨肌把被称为镫骨(以其形状来命名)的听骨从椭圆窗拉出来，而鼓膜张肌拉动附着在鼓膜上的槌状听骨，从而使槌状听骨变硬。结果就是声音被压制了。当我们说话时，镫骨肌通常处于紧张状态，来防止我们被自己的声音干扰。当我们吃东西的时候，鼓膜张肌也以同样的方式运作来减小我们自己咀嚼的音量。

声反射是引起**暂时性阈值偏移**(TTS)的原因之一，在这种情况下，我们认为比较安静的声音会变得完全听不见。不幸的是，要想使这种反射发挥作用需要一定的准备时间(称为延迟)，通常在45毫秒左右，这比枪声或爆炸声等冲动性声音造成的破坏所需的时间要长得多。反射作用只是TTS的一个因素，其他机制尚不清楚。

超负荷的耳朵与其他被滥用的(生物的和技术的)测量仪器的不同之处在于，重要的不仅仅是噪声的声压级，能量也很重要。持续一秒的噪声或许只会让人感到讨厌，但如果持续一小时就有可能导致严重的听力受损。TTS的数量与噪声出现时间的对数成正比，也就是说，时间延长一倍造成的效果远大于TTS的量翻一番。

然而，非常大的脉冲声，比如枪声，即使总能量很低也会造成瞬间的伤害。据说，一名训练有素的听力学家能够识别出开枪时没有戴耳罩的永久性听力损伤病人。恢复时间随音量的增大而增加，40分贝的TTS需要数周才能恢复。如果噪声中有一个停顿，TTS的数量就会大大减少。所以，如果必须长时间暴露在噪声中(比如在一场足球比赛中)，那么主动离开嘈杂的区域是最行之有效的方式，哪怕离开的时间只有短短15分钟。

根据大量关于噪声对听力影响的实验结果，许多国家的工作场所以及其他地方已经制定了噪声法规。它们对峰值噪声水平和暴露在指定较低水平下的时间进行了限制。

另一个问题是确定有多少人受到噪声影响导致听力损失。每个人听到高频声音的能力都会随年龄增长而下降。新生儿可以听到20千赫，大约40岁时降至16千赫，而60岁时下降到10千赫。80岁时，我们大多数人对8千赫以上的声音都是完全听不到的。这种现象被称为老年性听力障碍(字面意思为“老年聋”)。由于噪声引起的听力受损通常

也会影响更高的频率，而且随着年龄的增长，这种情况会越来越普遍，所以我们并不知道有多少损失是真正可以避免的。

在应对危险的高水平噪声的过程中，我们采取的行动往往很有效，如果一定要找出缺点来的话，应该是对那些需要足够响才能激怒人的噪声还不够重视，而且这类噪声的数量非常巨大。例如，2008(英国)国家噪声调查发现，有26%的受访者曾被邻居的噪声惹恼过。受影响的人睡眠不足，注意力不集中，失去耐心，失去生活的乐趣(39%的人说他们的生活质量受到了负面影响)。这种噪声还会加剧心理健康问题，导致社区紧张和社交孤立。

几乎没有人会否认，最好的方法是首先消除环境中的噪声。如何做到这一点，以及如何有效地做到这一点，关键在于控制声音来源。对大多数发达国家和许多其他地方的居民来说，下列因素可能是噪声的主要来源：

- 空中交通
- 工业
- 邻居
- 社区
- 铁路
- 道路交通

此外，船舶噪声和风电场噪声对一些地区也有较大影响，而大部分地区不时受到建筑噪声的影响。

在这些群体中，邻居和社区的噪声特别难以处理，部分原因是制造者没有把它们归类为噪声，对听者而言，它们却应该被归类为噪声。这种困难使得声音历史学家卡伦·柏斯特韦德(Karin Bijsterveld)得出“噪声控制被‘控制的悖论’所掌控”的结论。专

家和政客对于控制噪声的承诺实际上最终还是落到了邻居身上。他们也制定了非常复杂的应对政策来限制如飞机等其他形式的噪声，政策复杂到很少有人能接触或者理解，更别说批评了。

反击

除了部分友善的噪声和部分警告信号外，彻底消除噪声几乎是所有人的愿望。消除噪声能够实现吗？

在有些情况(不幸的是这种情况非常罕见)下，噪声是高度规则的，由一个或几个频率组成，并且只有在一个小而明确的位置上会引起麻烦。这时，**主动噪声消除**(ANC)是非常有效的方法。ANC的工作原理是声波由压缩波和稀疏波组成。在压力图(见图1)中，这一点很明显，因为这条线先高于中点，然后低于中点。如果在第一个声波上叠加第二个声波，使得在原声波受到压缩的每一点上，新声波都具有同样幅度的稀疏度，这样就会产生完全的静音，声波的能量就会完全转化为热能。然而，在实践中，这只能在非常有限的区域内实现，例如飞机驾驶舱、一些汽车的驾驶员头部区域，还有(可能最有用的应用)耳朵保护器的内部空间(见图23)。

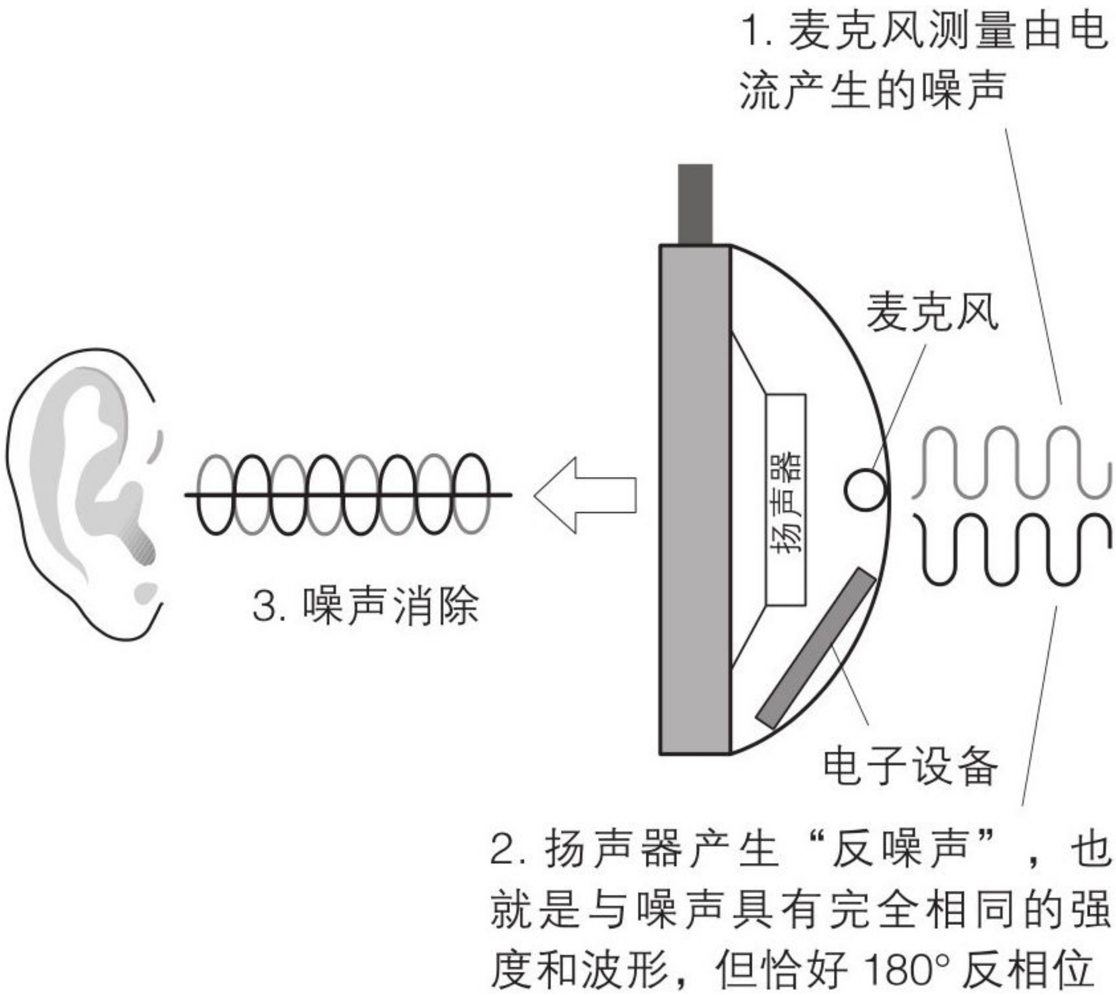


图23　主动噪声消除技术

我们已经开发出了许多针对噪声消除的高效技术解决方案，并且所有这些技术都已经被投入应用了。现代交通工具非常安静，比过去那些不能提供强大动力的交通工具相比要安静得多，但这一事实却被如今交通工具的巨大数量掩盖了。但是，我们仍然有必要看看今天的机器是如何变得如此安静的，因为同样的原理在未来也肯定会适用，尤其是在新材料出现的时候。

噪声控制的首要原则是识别和消除噪声源。对目前最重要的噪声源——**内燃机**(ICE)来说，这是不可能的。噪声是由内燃机内部的爆

炸产生的，而爆炸本身就有噪声。然而，正如第3章所解释的那样，我们的听觉系统并不能测量一个声音的总能量，它对4千赫左右的频率反应最强。让内燃机发出声音的频率低于这个频率是可以实现的，主要是调整燃料注入的时间和速度来改变燃料的燃烧过程，从而减少那些高频声音。

在处理了尽可能多的噪声源后，下一步是控制它。这方面，内燃机也很具挑战性，因为内燃机必须与周围的空气相连，吸入氧气并排出废气，而空气正是噪声传播的途径。为了减少噪声的排放，使用消声器是解决之道。枪支消声器其实就是这类东西，但只被用来降低枪击的声音。

消声器的基本原理有两个，大多数内燃机都会同时使用到两者。吸收性消声主要是通过在管道中加吸收剂来实现的，这样，任何不垂直向下传播的波都能转化为热量。那些垂直向下的波也减少了，因为管道内衬也削弱了这些波的能量。但是吸收性消声对低频噪声效果微弱，这时，用反应性消声器效果会更好，比如利用亥姆霍兹谐振器，将其调到最让人讨厌的声音的频率上。因此，这些频率的声音在消声器内部得到增强，它们的能量在那里被吸收。有一些消声器也使用主动噪声消除技术。

在接下来的几十年里，内燃机可能会在很大程度上被电动发动机所取代。从技术上来看，我们无法解释发动机为什么不能达到完全无声，但是一辆汽车在接近行人的过程中不发出声音，这种情况可能是致命的。所以我们会为所有这些场景提供特殊生成的运行和警告声音(这可能对自行车也有价值，尤其是那些没有铃铛的自行车)。然而，这些都不非得是让人不愉快的声音，即使警报声也可以不让人感到惊慌。为了吸引注意力，“嘘”声可以达到同样的效果，却没那么烦人。

另一方面，喷气发动机在可预见的未来可能还会继续存在。在喷气发动机中，噪声的来源并不是燃烧的气体本身，而是当它们从发动机中高速喷出后与相对缓慢的飞机外部空气混合时的效果，由此引起的湍流是几乎所有噪声的来源。为了使混合过程尽可能平缓，人们付出了巨大的努力，让射流产生波纹，在中速空气中形成茧状，从而增大射流的圆周。在这种设计思路上再进一步就不得不严重损失推力了。但也有其他方法可以考虑。有一个想法已取得了一定的成功，即在机翼上方安装发动机。另一个方法是战略上的改变，比如使用飞艇来运输那些并不紧急的货物。

除最慢或最老式的车辆外，所有的车辆都采用了流线型设计，这样既能避免产生噪声，又能提高效率。流线型设计的重点一方面在于找出一个合适的形状，使气流滑过它时不会突然改变方向或速度而导致湍流；另一方面在于制造出光滑的表面。然而，就汽车驾驶员而言，流线型所要处理的空气动力噪声在巡航和高速行驶时仍然是不可避免的。主要的声音来源通常是后视镜和A柱(那些支撑风挡玻璃两侧的柱子)。

对人行道上的行人来说，最大的噪声往往是由轮胎发出的，尤其是在混凝土路面上。低噪声的表面可以在很大程度上解决这个问题。目前可用的最好的沥青层有许多充满空气的孔隙(约占体积的25%)，但这样的路面成本高昂，而且不够坚固。另一种效果较差，但更耐用、成本更低的替代方法是使用约2厘米厚的薄沥青层。

曾经，机器的设计者根本不考虑它是否安静，噪声管理只是作为事后的一种补救或不可靠的常规做法(比如在音乐厅地板下拉几根绳子，或在剧院舞台上摆放花瓶)。现在，机器通常经过精心设计和严格制造，能够保持相对安静，不过，工业机械的定期维护和家用电器的仔细安装仍然必不可少。

当噪声无法避免而又无法控制时，下一步就是把它的来源与潜在的受害者隔离开来。一种使用了数千年的方法是分区制。通过立法将产生噪声的活动限制在特定区域，例如远离居民区的工业区。第一个有记录的例子可以追溯到公元前700年左右，当时在爱琴海沿岸的希腊殖民地，锡巴里斯市中心禁止铁匠、木匠、陶工，甚至公鸡进入。

由于历史或地理因素使得分区隔离行不通的时候，设置声音屏障是主要的解决方案。一个只是让人们看不到噪声源的障碍就能减少5分贝的噪声水平，在此基础上，距离每多一米将额外减少1.5分贝。此外，屏障的密度必须足够大，比如每平方米10千克左右就足够了，这样就可以防止很多声音直接穿过它。由于屏障在很大程度上是反射而不是吸收，因此我们还要考虑反射的声音。除此之外，屏障的设计和施工都很简单，结果是可以预测的，成本也相对较低。

一些临时噪声隔离屏障，即隔声屏障(acoustic fencing)十分灵巧，可放置在建筑地盘的四周。小心放置、周全安排，不留下空隙的话，隔音效果将会非常好，在某些情况下减低的噪声可达30分贝。与永久性隔声屏障不同，临时性隔声屏障主要通过吸收来实现降噪。

隔声屏障最大的问题是它们并不让人觉得赏心悦目。这个问题可以通过在屏障和观察者之间引入植被屏障来抵消。如果在噪声和人群之间种植至少10米宽的植被，也有助于减少噪声。低频噪声(约25赫兹)主要是由铺满落叶的地面承担降噪工作，而高频噪声(超过1千赫)主要是通过树叶降噪，树叶的长度是声波长度的一半时效果最好。由于这两个原因，种植一些落叶植物(树和灌木的混合)是必不可少的，而且应当达到尽可能高的高度(高达10米，比这更高的话，从大树枝反射的效果则会降低)。在有风的日子里，植物的簌簌声可以提供另一个额外的好处，即分散人们对噪声的注意力。

在城市公园中，添加自然声源(声景)的方法尤其有效。在这样的空间中，最常见的改善声景的方法是添加喷泉，鸟鸣声(真实的或录制的)也有不错的效果。

安静的家居环境

在所有的空间中，最珍贵的是家，因为家是我们放松和睡觉的地方，同时它也是我们个人空间的延伸。因此，消除居家噪声是控制噪声的首要任务。最好的方法是进行良好的声学设计，包括选择合适的材料，以及正确地安装和维护。对已经建成的住宅来说，良好的声学设计作为改造噪声的方案，更便宜、更美观，也更容易成功。

无论改造与否，减少家中噪声的基本方法很简单，一是阻止噪声进入，二是破坏已经进入的噪声，三是不要再制造更多噪声。声音有三种进入方式：通过入口传播；通过结构传递振动；以及将墙壁、窗户、门、天花板和地板作为通道进行传播。在这三种情况下，我们需要记住的主要一点是外壳的声学性能取决于它最薄弱的部分，就像船体上的一个小孔会使船的水密性能失效一样，双层玻璃房子上一扇开着的窗户也是如此。事实上，由于我们耳朵的对数反应[1]机制，有噪声的情况比有水的情况要糟糕得多。如果我们将船上两个相同的孔中的一个封住，水量将减半。但如果我们把一所房子上两扇相同的窗户中的一扇关上，进入的声音也会减少一半，可是那50%的声强下降所对应的，只是大约2%的响度下降。

防止噪声的第二种方法是双层玻璃窗，因为两个单层玻璃窗构成了很好的分隔。通过结构传播的声音是一个更大的挑战。如果声源在住宅内部，那么可以把那个房间处理成不利于声音传播的结构。但这种改造并不便宜，而且通常要对天花板、所有墙壁、地板及门做同样的处理。

一种廉价、适应性强且有效的解决方案是悬挂带有尽可能多褶皱的厚重丝绒窗帘，但这可能不是最吸引人的方案。如果需要更极致的方案的话，就必须得有专家参与解决了。还有一个显而易见的解决方案是加厚墙壁，要记住，墙体厚度每翻倍一次，声音的损耗只有6分贝(声功率降低约3/4，但音量只减少40%)，这意味着墙要非常厚才能达到理想的预期隔音效果。

一个更好的方法是使用多孔吸收体和多层结构。像玻璃纤维这样的多孔吸收体，高频声波在其中传播时要多次通过其表面，会逐渐因反射而消失。低频波在空气或固体界面反射(声阻抗不匹配的另一个例子)，也能用带气隙的嵌板和墙体处理。一个合适的隔音门也是至关重要的。

地板也不容忽视。即使下面没有房间，硬地板在产生走路时的噪声和在整个建筑中噪声的传递方面都很惊人。地毯和衬底对高频声波的隔离非常有效，但在隔离低频声波时几乎毫无用处，因此，如果你家里有一个安装在地板上的扬声器或一个外向的男中音，你就需要一些特别的东西。和之前一样，除了请一位专家帮忙，其实也没有其他更好的选择。

虽然在家里回声很少引起困扰，但公共空间设计中最大的问题就是回声。令人惊讶的是，数千年来，尽管对建筑师、表演者和观众来说，这都是一个棘手的难题，但在1898年之前我们都没有办法量化它。1898年，华莱士·克莱门特·萨宾(Wallace Clement Sabine)将房间的回声时间定义为声强降至初始水平的百万分之一(-60分贝)所需的时间(见框13)。更有用的计算公式是萨宾推导出的一个基于经验的方程，这个方程允许从房间的大小和形状来计算这个时间，这样建筑师就可以预测回声时间了。

框 13　萨宾方程

$$T=0.161(V/A)$$

其中，T 为回声时间，V 为房间体积，A 为总吸收面积。除了高度消声的房间外，这个公式如今还在广泛使用着。

除非事先对问题的根源和周围的环境进行仔细调查，否则在一个特定的噪声问题上无论花费多少心思、金钱和专业知识，结果都将是无效的。如果门常常是敞开的状态，那么即使将普通的门换成隔音门也没用，噪声问题不会被解决。如果车流和人流会通过居民区的街道，那么仅仅对路上行驶车的辆进行限速，这对消减噪声也没什么用。这样整体降噪产生的问题远远超过消除噪声这个问题本身。安装吸声板能够带来的潜在益处必须能与它们同时会造成的视觉冲击相平衡。同样地，试图通过引入人工声音分散公园中的人们对交通噪声的注意，可能会造成更大的反感。把人从开放式办公室转到安静的小隔间可能会降低团队的工作效率。同样，虽然建造良好的房屋可以隔绝噪声，但这必然导致把居民关在里面，也隔绝了自然的声音，产生人际异化。正如戴维·亨迪(David Hendy)所说：“每当我们退回各自的声音场景……我们彼此成了陌生人。”

更广泛的观点

尽管听觉系统非常重要，但它们的性能对我们而言通常不是生死攸关的问题。不过，对有些水下物种来说，听觉系统却关乎存亡。对

许多海洋动物来说声音极为重要，声音在水下的传播意味着一些海洋生物会因噪声的影响而陷入危险，尤其是鲸鱼和海豚。这种影响的范围尚不清楚，但有些影响是极其微妙的。由于鲸鱼通过声音信号与幼鲸交流，母子之间的联系很容易被外来的声音打断，这会导致幼鲸与母亲分开。在另一些情况下，巨大的不熟悉的声音对鲸鱼和海豚的惊吓作用非常之极端，导致它们快速浮出水面，肝脏和其他器官中形成氮气气泡而死亡(潜水员将这种情况称为“减压病”)。

水下噪声污染的成因有很多。除了船舶的声音，声呐系统、爆破和通讯信号都增加了水下生物的负荷。幸运的是，现在人们普遍认识到噪声对动物的影响。事实上，这种情况对潜水者而言危害更大，他们经常面临严重听力损伤的风险(由于这种影响往往与耳朵压力变化的影响相同，因此很难确定问题的严重程度)。

我们在陆地和水下与噪声做的斗争具有跨学科属性和公益性质，所以社会各层的合作至关重要。如果没有世界卫生组织等国际组织、各国政府、地方当局、英国噪声控制协会等专业组织以及公众的积极支持，我们就无法取得成功。能够把所有这些参与者聚集在一起的活动有“噪声行动周”和“国际噪音意识日”。

随着技术的进步，城市的发展，汽车旅行的普及，噪声源的增加，噪声污染已经成为许多人生活中不可避免的问题：无论是通过扬声器有意地、自私地制造的噪声，还是通过发动机和其他硬件不可避免地、无意地制造的噪声。科学可以尽可能地降低噪声，但去除噪声污染必须伴随着教育和立法，以确保噪声被大众视为污染物。

这样，制造噪声对大众而言将不仅仅是恼人而已，它将变得不可接受。虽然这似乎不切实际，但英国在禁烟方面确实发生了与之类似的变化，而且只花了几年时间。这一变化是通过国家和地方政府实施的一系列新法律，以及一个响亮而明确的信息——吸烟会导致严重疾

病——促成的。如果我们用同样的方法处理噪声，也能产生彻底的改变。

在所有学科中，与声音相关的受众群体最广泛。通过了解声音，演员、广告商、航空航天工程师、汽车工程师、人类学家、建筑师、艺术家、广播员、建筑商、通信工程师、作曲家、设计师、生态学家、教育家、电子工程师、环境卫生官员、电影制作人、历史学家、海洋生物学家、音乐家、医生、物理学家、勘探者、心理学家、地震学家、社会学家、城市规划者、动物学家和许多其他人的工作效率将得以提升。

到目前为止，声学领域的许多方面都是各自独立的。但是在过去的几十年里，像声音采集、声音研究以及设计更加私密和个性化的媒体系统等方法已经开始汇集来自各个领域的专业知识。这种合作往往受到大多数从业者观念狭隘的阻碍，但与此同时，他们也正在让人们更加清楚地意识到这个主题涉及的广阔范围，以及声音本身所拥有的权利。掌握声音的原理能够带来大量实际的和情感的利益，将声音科学的众多分支交叉融合起来也能够创造许多新的应用场景，并产生新的认知。这样的发展对每个人都有利，对我们所有人来说，未来都值得期待。

全书完

[1] 非线性变化，人耳对响度和频率的实际感受更接近对数变化。

术语对照

A

A weighting　A权重

absolute pitch　绝对音准(也叫绝对音高)

acoustic fencing　隔声屏障

acoustic impedance　声阻抗

acoustic lens　声透镜

acoustic microscope　声学显微镜

Acoustical Thermometry of Ocean Climate　海洋气候声学测温，ATOC

acousto-optic effect　声光效应

active loudspeaker　有源扬声器

active noise cancellation　主动噪声消除，ANC

aircraft malus　飞机惩罚

antiphase　反相位

attack sound　突发声音

Attali, Jacques　雅克·阿塔利

audiology　听力学

auditory canal　耳道

auditory nerve　听觉神经

auto-changer　自动换片器

automatic sound recognition　自动声音识别技术

automatic speech recognition　自动语音识别

B

backwave　反射回波

baffle (of loudspeaker)　(扬声器的)挡板

bar　小节(音乐术语)

Barthes, Roland　罗兰·巴特

basilar membrane　基底膜

bass　低音

beat　节拍

Bell, Alexander Graham　亚历山大·格雷厄姆·贝尔

Bijsterveld, Karin　卡伦·柏斯特韦德

birdsong　鸟鸣

Bloop　海洋怪声

body wave　人浪

bolide　火流星

bone conduction　骨传导

breaking of voice　变声

Broca’s area　布罗卡氏区(大脑结构)

Bull, Michael　迈克尔·布尔

C

cassette recorder　盒式录音机

cathode ray oscilloscope　阴极射线示波器，CRO

cavitation　气蚀

cent　音分(音乐术语)

Chesapeake Bay　切萨皮克湾

Chladni, Ernest　恩斯特·切尔德尼

chord　和弦（音乐术语）

cochlea　耳蜗

cochlear implant　耳蜗植入

cocktail party effect 鸡尾酒会效应

Comprehensive Nuclear-Test-Ban Treaty Organization 全面禁止核试验条约组织，CTBTO

consonance 和谐

consonant 辅音

constructive interference 相长干涉

Corbin, Alain 阿兰·科尔宾

cymbal 铙钹

D

Darwin, Charles 查尔斯·达尔文

deafness 耳聋

decibel 分贝

deep sound channel 深海声道

destructive interference 相消干涉

difference frequency 差频

diffraction of diffraction grating 衍射光栅

Dissonance Quartet (Mozart) 《弦乐四重奏》(莫扎特)

distortion (of ultrasound) (超声)失真

Dolby system 杜比系统

Doppler effect 多普勒效应

E

ear defenders 耳朵保护器

ear phonautograph 人耳记音器

eardrum 鼓膜

earworm 耳虫

echolocation 回声定位

Edison, Thomas 托马斯·爱迪生

electronic music 电子音乐

end correction 末端校正

Eno, Brian 布莱恩·伊诺

equal temperament 十二平均律(音乐术语)

eustachian tube 咽鼓管

F

fast Fourier transform 快速傅里叶变换，FFT

filter 滤波器

flank array　侧翼阵

flat　降调(音乐术语)

fluctuation　波动

focussing array　聚焦阵列

foetal scanning　胎儿扫描

Foley　拟音

formant　共振峰

Fourier analysis　傅里叶分析

Fourier series　傅里叶级数

Fourier transform　傅里叶变换

Fourier， Joseph　约瑟夫·傅里叶

frequency sweep　扫频器

fricative　摩擦音

fundamental frequency (vocal)　(人声的)基频

futurism　未来主义

G

Galilei, Vincenzio　文森佐·伽利雷

Garland, Judy 朱迪`加兰

geophone 地音探听器

glissando 滑音

Gran Concerto Futuristico (Russolo) 《 未来大协奏曲》(鲁索洛)

H

Haas effect 哈斯效应

hair cell 毛细胞

harmonic series 和谐音程

hearing aid 助听器

hearing loss 听力损失

Helmholtz resonance 亥姆霍兹谐振

Helmholtz resonator frequency 亥姆霍兹谐振器频率

Hendy, David 戴维·亨迪

High-intensity focussed ultrasound 高强度聚焦超声，HIFU

High-intensity therapeutic ultrasound 高强度治疗超声，HITU

Hooke, Robert 胡克·罗伯特

Huffman coding 霍夫曼编码

Hughes, David　大卫·休斯

I

impulse　脉冲声

infrasound　次声

K

karaoke　卡拉OK

Katz, Mark　马克·卡茨

key　调(音乐术语)

key note　主调

Kircher, Athanasius　亚森纳希思·柯歇尔

L

Labelle, Brandon　布兰登·拉贝尔

Laplace, Pierre Simon　皮埃尔·西蒙·拉普拉斯

lateral line　侧线

law of the first wavefront　第一波前定律

Leonardo da Vinci　列奥纳多·达·芬奇

liquid　流音(辅音的一种)

London Eye　伦敦眼

Long Range Acoustic Device　远程声学设备，LRAD

longitudinal wave　纵波

Long-playing record　密纹唱片，LP

loudspeaker　扬声器

Lully, Jean-Baptiste　让-巴普提斯特·卢利

M

major scale　大调音阶

mastoid bone　乳突骨

measurement speaker　测量麦克风

MEMS speaker　微电子机械系统麦克风

microelectromechanical systems　微电子机械系统 MEMS

microtone　微分音

middle C　中央C

Miller, George　乔治·米勒

minor　小调

monochord　单弦琴

Morland, Samuel　萨缪尔·莫兰

Mozart, Wolfgang　沃尔夫冈·莫扎特

MP3 player　MP3播放器

MRI scan　核磁共振扫描

muzak　助兴音乐

N

Neanderthals　尼安德特人

neuromast　神经丘

Newton, Isaac　艾萨克·牛顿

Newton’s cradle　牛顿摆

Noise Abatement Society　英国噪声控制协会

Nyquist theorem　奈奎斯特定理

O

octave　八度

organ of Corti　柯蒂氏器

ossicle　听小骨

otoacoustic emission　耳声发射

otolith　耳石

oval window 卵圆窗(耳部构造)

overtone 泛音

P

Pacinian corpuscle 帕西尼安小体

paradox of control 控制的悖论

parametric array 参数阵列

pentatonic scale 五声音阶

persistence of vision 视觉暂留现象

Philip, Robert 罗伯特·菲利普

Philipsz, Susan 苏珊·菲利普斯

phonons 声子

phonograph 留声机

pickup 拾音器

piezoelectric effect 压电效应

Pinker, Steven 史蒂芬·平克

pitch 音高

plosive 爆破音

polyvinylidene bifluoride　聚偏氟乙烯，PVDF

presbyacusis　老年聋

primary auditory area　主听觉区

psychoacoustics　心理声学

pure tone　纯音

Pythagoras　毕达哥拉斯

Q

quantum mechanics　量子力学

R

railway bouns　铁路奖励

reciprocal transducer　互易换能器

rectifier　整流器

reference pressure　参考压力

resonance　共振

rests　休止符(音乐术语)

reverberation　混响

rhythm　节奏

Rite of Spring (Stravinsky) 《 春之祭》(斯特拉文斯基)

round window 圆窗(耳部构造)

Russolo, Luigi 路易吉·鲁索洛

S

Sabine, Wallace Clement 华莱士·克莱门特·萨宾

Sacks, Oliver 奥利弗·萨克斯

Schafer, Murray 穆瑞·谢弗

seismic wave 地震波

semibreve 全音符(音乐术语)

semi-circular canal 半规管(耳部构造)

semitone 半音

semivowel 半元音

sharp 升号(音乐术语)

Shepard tone 谢泼德音调

signal to noise ratio 信噪比

Snell’s law 斯涅尔定律

sonic boom 音爆

sonobuoy　声呐浮标

sonochemistry　声化学

sonoluminescence　声致发光

sound fixing and ranging　声音定位和测距，SOFAR

sound pressure level　声压级，SPL

soundscape　声景

spectrogram　声谱图

spherical spreading　球面传播

spiral ganglion neuron　螺旋神经节神经元

square wave　方波

standing (stationary) wave　驻波

stapedius muscle　镫骨肌(耳部构造)

stem cell　干细胞

stereocilia　静纤毛

Sterne, Jonathan　乔纳森·斯特恩

Stokes' law　斯托克斯定律

Stravinsky, Igor　伊戈尔·斯特拉文斯基

structural health monitoring 结构健康监测

subharmonic 次谐波

supramarginal gyrus 缘上回(大脑结构)

Szendy, Peter 彼得·森迪

T

temporary threshold shift 暂时性阈值偏移，TTS

tensor tympani muscle 鼓膜张肌

Thompson, Emily 艾米丽·汤普森

threshold of hearing 听力阈值

threshold of pain 疼痛阈值

Tonpilz transducer 蘑菇形压电换能器

trachea 气管

transducer 换能器

transistor 晶体管

transverse wave 横波

travelling (progressive) wave 行波

treble 高音

triode　三极管

tuning fork　音叉

Turner Prize　特纳奖

tympanic membrane　鼓膜

U

ultrasound　超声

underground nuclear tests　地下核试验

underground sound　地下声音

undertone　底音

unvoiced consonants　清辅音

V

velocity of sound　声速

vibrato　颤音

vocal cavity　声腔

vocal folds (vocal cords)　声带

vocal tract　声道

voice coding　声音编码

voice coil　音圈

voice prints　声学指纹

voiced consonants　浊辅音

voix cèleste　天籁音栓

W

Wernicke’s area　韦尼克区（大脑结构）

Wheatstone, Charles　查尔斯·惠斯通

whip crack　响鞭

whispering gallery　耳语廊

white noise　白噪声

White, Graham　格雷汉姆·怀特

White, Shane　肖恩·怀特

World Forum for Acoustic Ecology　世界声学生态论坛

World Health Organization　世界卫生组织

World Soundscape Project　世界声景计划

X

xenon　氙气

Z

zoning　分区制

Mike Goldsmith

SOUND

A Very Short Introduction

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© Mike Goldsmith 2015

The moral rights of the author have been asserted

First edition published in 2015

Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2015945427

ISBN 978–0–19–870844–5

Printed in Great Britain by
Ashford Colour Press Ltd, Gosport, Hampshire

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

With thanks to

Richard Barham

Graham Carter

Mark Hodnett

Stephen Robinson

Hilary Whittle

Contents

List of illustrations

1 Sound wave pressure plot 11

2 Molecular view of sound wave 12

3 Sound propagation when air is cooler near the ground than high above it 14

4 Barbara can hear Alan, but not Clive 14

5 Difiraction 17

6 Difiraction of difierent wavelengths 18

7 The acousto-optic efiect 19

8 Constructive and destructive interference 19

9 Wave shapes 22

10 Standing waves in a pipe open at one end 28

11 Spectrogram 31

Raphael Frey/Wikipedia commons/Creative Commons AttributionShare Alike 3.0 Unported

12 Summing sine waves to approximate a square wave 32

13 Beats 48

Joe Wolfe/〈 http://www.animations.physics.unsw.edu.au/jw/beats.htm〉

14 The ear (the middle and inner ears are greatly enlarged) 53

Figure 2a, ‘Mechanism of the Ear’, in Mike Goldsmith, Discord: The Story of Noise (Oxford University Press, 2014), p. 10. By permission of Oxford University Press

15 Frequency dependence of hearing 54

16 Locations of hearing, language, and speech activities in the brain 59

17 Vocal apparatus 65

18 Dynamic microphone 70

19 Condenser microphone 71

20 Loudspeaker 74

21 Ripples vs waves 84

22 Underwater sound velocity proffle 103

Xavier Lurton, An Introduction to Underwater Acoustics: Principles and Applications (Springer Praxis, 2010), p. 40. With permission of Springer Science, Business Media.

23 Active noise cancellation 115

The sound spectrum

	Frequency in hertz
Low-frequency volcanic activity	0.01
Airborne infrasound from sea waves	0.20
Lower limit of elephant hearing	16
Minimum audible tone	20
Lowest piano key	27.5
Lowest recorded sung note	41
Middle C	261.626
Piano key A above middle C	440.000
A above middle C on Bach's organ	480
Highest recorded sung note	1,865
First part of blackbird song	2,500
Highest piano key	4,186.01
highest audible tone, aged 80	8,000
Highest audible tone, aged 10	20,000
Dental cleaning	25,000
Typical ultrasonic welder	30,000
Ultrasonic parking sensor	40,000
Upper limit of dog hearing	60,000
Typical bat scan frequency	80,000
Upper limit of bat hearing	110,000
Upper limit of beluga whale hearing	123,000
Typical fetal scan	15,000,000
Acoustic microscope	8,000,000,000

Chapter 1 Past sounds

Sound: 13.7 billion years ago

Sound has its origin far back in time, not long after the disappointingly silent Big Bang. In fact, sound waves formed as soon as there was a medium for them, which was 300,000 years after the beginning of everything, and thirteen billion years before there was anyone to listen.

The primordial sound was of a very low frequency, but powerful and omnipresent, and it formed when the plasma of the newborn universe arranged itself in pseudo-regular patterns through space. Galaxies eventually formed in the denser regions, including the progenitor of the world we live on and the sun we orbit.

Moving on a few billion years to the first days of Earth (about4.6 billion years ago), there were sounds aplenty, passing through the solid planet's crust and its subsurface liquid areas, and bouncing and bending through its atmosphere. Eventually, the hot land cooled, rains fell, and oceans formed—oceans full of sound. So, the environment in which the first living things evolved was an acoustically rich one, which profoundly affected the forms, habits, and destinies of those creatures.

Hearing: 500 million years ago

For us, there is a clear distinction between hearing and feeling, but it's a very different matter for undersea creatures: sounds pass as easily through their bodies as vibrations do through ours, and, in fish, are detected by structures called neuromats, which are distributed over their body surface (fish have several other hearing structures too).

Neuromats contain hair cells similar to those in our ears and provide their owner with information about the strengths and directions of local sounds. They evolved perhaps 500 million years ago. To detect airborne sound, eardrums and cochleae are essential, and hence evolved once amphibians began to colonize the land, around 400 million years ago. Communication was probably the main spur to the evolution of hearing, since sounds have overwhelming advantages over visual signals: while the ability to make lights and change colours exists in some marine organisms, putting on a light show is far more challenging and narrower in range than making noise. Noises are made easily—as easily as breathing. In humans, breath-made sounds (precisely controlled by our big brains) gave us the power of speech.

Music: 40,000 years ago

The appreciation of music is a mysterious pleasure and an ancient one too: over 40,000 years ago Neanderthals probably had flutes and, by the time Homo sapiens emerged, no well-appointed cave dwelling was complete without a rock gong. It

may be that the first humans sang; perhaps even before speech evolved. But why? Enjoying music has no obvious evolutionary advantage. Darwin himself was baffled, but suggested that a taste for music might arise through sounds made in mating rituals, and many echo this view today. Others, however, prefer the suggestion of evolutionary psychologist Steven Pinker that music is an auditory equivalent of cheesecake, which we enjoy not because that preference assisted our ancestors in surviving, but because many of the sensations cheesecake elicits are of evolutionary value in themselves: the sweetness of fruit announces its ripeness and creamy flavours suggest energy-rich fats. Or perhaps music reminds us of birdsong—the presence of which indicates that no large predators are around.

Harmony: 2,500 years ago

Today, sound plays a huge variety of roles in our lives. Many of our inventions are dedicated to its creation, transmission, storage, modification, or reproduction. But the conquest of sound is by no means a recent development: some of the most ancient artefacts we know are musical instruments, and acoustics was one of the first sciences; in 500 bce or so, Pythagoras discovered that the sound made by a strummed string mingles pleasantly with one made when the string's length is halved. The 'distance' between the two sounds is an octave, by definition and by universal agreement the most harmonious of all pairs of different notes. Sounds almost as harmonious result if the string lengths bear other

simple numerical ratios: if one string is one-and-a-half times the other, a fifth is produced, for example.

According to legend, Pythagoras made this discovery when he heard tuneful hammer sounds emerging from a forge where a number of blacksmiths were at work. When (being a budding scientist) he weighed the hammers they were using, he found that those pairs which made a pleasant sound had weights which were simple multiples of each other. The fact that this story is told to this day is surprising given that the frequency at which a hammer sounds is not fixed by its weight. Whatever actually piqued his interest, the instrument Pythagoras used to study harmony was the monochord—a device with a single string whose length can be set by a moveable bridge.

To Pythagoras, the fact that the pleasantness of sounds was defined through whole-number ratios suggested that numbers were the key to the universe. 'All', so he is supposed to have said, 'is number'. Today's scientists would agree, and, in its impact on scientific method, mathematics, music making, and acoustics, his discovery may be one of the greatest breakthroughs of all.

Although anyone who makes or plays a stringed instrument knows that tension, as well as length, affects the note a string makes (otherwise turning pegs to tune stringed instruments wouldn't work), this was not quantified until the 16th century by Vincenzo Galileo, father of the

scientist, who showed that pitch increases with the square root of the tension. We now know that it also depends on the string’s thickness and density.

The Greeks were interested in the practicalities of sound, thanks to their keen interest in making their voices heard: plays, orations, declamations, debates, songs, chants, and proclamations abounded. Perhaps their greatest acoustic structure is the theatre of Epidaurus, built in the 4th century bce. The distance from stage centre to the back row is about 60 metres, yet actors can be heard clearly from any of the 1,400 seats: fifty-five rows in total. It is in these seats that the theatre’s acoustic secret lies: the limestone of which they are made, their corrugated surfaces, and the spaces between them all contribute to absorbing sub-500 Hz frequencies and reflecting higher ones, quietening crowd murmur and enhancing performances respectively.

But there are disadvantages to open air speech, even in Epidaurus. With no ceiling to contain the sound, speakers must be very loud indeed, which is not only tiring but also tends to rob the voice of its subtlety. (Despite legends to the contrary, the Greeks lacked megaphones, which were not invented until the 1670s—simultaneously by Athanasius Kircher in Germany and Samuel Morland in England.) Background noise becomes far more intrusive too, though in the Greek theatres performance days would have been quiet, since the majority of the local population would be in the theatre.

Indoor public spaces solved these problems but introduced new ones: echo and reverberation. For an echo to be an echo, it must be heard more than about 1/20 of a second after the sound itself. If heard before that, the ear responds as if to a single, louder, sound. Thus 1/20 second is the auditory equivalent to the 1/5 of a second that our eyes need to see a changing thing as two separate images. (Hence, when camera frames move faster than this, we obtain the illusion of flowing movement. This ‘persistence of vision’ is what makes us see the rapid sequence of still images that make up a cinema film as a smoothly changing image.)

Since airborne sounds travel about 10 metres in 1/20 second, rooms larger than this (in any dimension) are echo chambers waiting to happen. Luckily echo can be reduced by covering hard surfaces with soft, fabric-covered objects, such as audience members.

There is of course more to sound than science or entertainment. Even wordless sounds come freighted with meaning, much of which was attached in prehistoric times. The lonely howl of the wind, a shocking scream of pain, the joyous songs of birds, the happiness of children’s laughter: in these and many other cases, evolution has forged unbreakable links between sound and emotion. These emotional attachments have been exploited by us since ancient times: war cries, for example, have long been produced both to chill the blood of the enemy and to unite and rouse the courage of the attackers.

The modern world: acoustics and more

Following the work of the ancient Greeks, little research on the nature of sound was carried out until the 17th century, when Robert Hooke proved by simple demonstrations that frequency and pitch are linked. Although Isaac Newton suggested an equation for the velocity of sound, it was incorrect, and an accurate version was only derived by Pierre Simon Laplace in

1816. Laplace showed that, to an excellent approximation, the velocity of sound depends only on the density and the elasticity of the medium through which it passes (Box 1).

Box 1

Frequency of a stretched string $\frac{1}{2l}\sqrt{\frac{T}{\mu}}$; length l, tension T, density μ (mu).

Velocity of sound v (or occasionally c from the Latin *celeritas*) = $\sqrt{K/\rho}$; elasticity K, density fi (rho).

The invention of the first electroacoustic devices in the mid-19th century led to revolutions in both the understanding and the control of sound: the microphone, telephone, and loudspeaker appeared in quick succession, spurring rapid developments in research, commerce, and the arts.

The 20th century and electronic engineering began together, with the inventions of the diode (the first rectifier, used initially for detecting radio signals) in 1903 and the triode (the first amplifier) in 1906. The development of electronics was greatly accelerated by the

World Wars, which also led to the birth of underwater acoustics research through interest in submarine warfare and ship detection.

While the existence of sounds with frequencies too high to hear had occasionally been discussed in the 19th century, they were only investigated in the context of the upper frequency limit of human hearing. Even though such sounds could be made easily by sparks, whistles, air jets, or piezoelectric crystals, they remained of little interest until World War I, when it was realized that they might be pressed into service as part of what we would now call sonar systems. Soon after the war, further investigations revealed that such sounds had a range of unique properties, not all of which could be explained: they could kill living things, cause chemical changes, generate light and heat, and make wood explode in showers of sparks. It may be that they were called supersounds in part because of their strange powers. Although far less famous than X-rays and radium emanations, supersounds soon gained a similar reputation as secret forces wrested from nature by the tools of science, but still retaining a glamour of almost supernatural mystery and power.

It was the middle of the 20th century before the power of ultrasound was properly understood and exploited. By then, the use of amplifiable electroacoustic technology had truly transformed the world. Public speakers, formerly limited to audiences of a couple of thousand at the very most by the sounds of their voices and the sizes of their venues, could

now talk to people in their millions, thousands of miles away—either instantaneously or a day or a century later. Life could be captured, recorded, and analysed as never before.

There were whole new fields too, including sound art; an ill-defined discipline with origins in the futurist movement of the 1900s to 1930s, and in the development of electronic music and good recording technology. Luigi Russolo's *Gran Concerto Futuristico* (1917) is an important early example, and a recent one is *Lowlands* by Susan Philipsz, a set of variations on a lament played over one another, which won the Turner Prize in 2010. A related area is ambient music (often called muzak by its detractors), which is designed to provide an appropriate background to public spaces. Brian Eno's *Ambient 1: Music for Airports* (1978) is a good example, and saccharine carols looped endlessly in supermarkets (accompanied by gloomy employees forced to dress as elves) is a bad one.

Ambient music is an example of an artificial soundscape, a concept popularized by Murray Schafer, who helped set up the World Soundscape Project in Vancouver in the late 1960s. The project led in turn to the formation of the World Forum for Acoustic Ecology in 1993. Schafer has been highly influential, in part through his galvanizing claim that acoustic environments can not only reveal the social conditions of those who inhabit them, but even predict how that society will evolve.

Applying Schafer's approach, economist and polymath Jacques Attali argues that changes in musical convention prefigure wider changes in society. Some have gone much farther than this; historian Alain Corbin argues that village bell ringing in 18th- and 19th-century French villages moulded social and economic relations there, and artist and writer Brandon Labelle says that 'my feeling is that an entire history and culture can be found within a single sound'.

More generally, the concept of the soundscape has enjoyed great popularity in a range of disciplines, though Schafer's definition of the term has been extended to take account of the relative and dynamic nature of an acoustic environment: technology historian Emily Thompson points out that it is 'simultaneously a physical environment and a way of perceiving that environment'.

In film, the construction of artificial soundscapes is achieved partly through sound effects. These have also been a mainstay of radio drama since its inception and, in the form of artificial thunder for example, have been heard in theatres since ancient Greece. In films, the craft of designing, producing, and synchronizing sound effects with events on screen is known as Foley.

No longer dependent on our ears to detect sounds or on voices and mechanical devices to make them, we can study and use sounds too low pitched, high pitched, or quiet to hear,

and we can generate and direct acoustic beams with enormous power and precision, leading to applications in medicine, defence, mapping, and many other fields. Since World War II, serious attempts have been made to develop sound-based weapons, mainly by producing and directing extremely high-intensity beams. Perhaps the best-known example in current use is the Long Range Acoustic Device (LRAD), which projects either commands or unpleasant sounds. It has been used against humans and wildlife in several countries.

Alongside the deliberate and controlled development of sound, noise pollution has also spread through much of the world. The highly sensitive hearing systems that served our ancestors so well, and still allow our pleasure in music and our facility in speech, are now the conduits of annoyance, stress, and damage.

So while we have gained mastery of the production of sound, we are far from being able to control it. In order to have any chance of doing so, we need to understand its nature.

Chapter 2 The nature of sound

Two faces of sound

If a tree falls in a forest without anyone to hear it, does it make a sound? The dual meaning of ‘sound’, as physical phenomenon and sensation, provides a clear answer: yes and no. The relationships between the physical and sensual aspects of sound are complex, in that many of the impressions sound makes on us are related to its physical parameters but not reducible to them. So: high-frequency sounds sound higher pitched—usually. And more powerful sounds sound louder—on the whole. Furthermore, many sounds, from sirens to skirls and from lullabies to lions’ roars, make emotional impacts on us which have only the vaguest relationships to their physical parameters.

The physical aspects of sound are far better understood than the emotional ones, so it is with physics that we should begin.

Pressure waves

Sounds are usually made by something moving in a cyclic manner: the diaphragm of a loudspeaker pulsing in and out, the gap between the vocal folds narrowing and widening, or a guitar string vibrating back and forth. It is the transmission of these motions to the surrounding medium (solid, liquid, or gas), and their progression through that

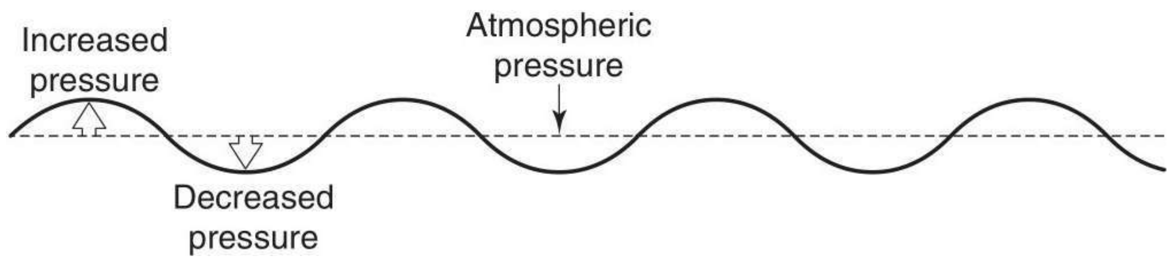
medium, that constitute sound. In some cases the motion begins in the medium itself, such as the air in the neck of a bottle when one blows across it. Non-moving sources include sudden releases of heat energy, such as by explosions or sparks, and rapidly oscillating heat sources.

When the motion is that of a loudspeaker’s diaphragm, the cause is a varying electrical signal with the same pattern as the sound wave that the diaphragm will produce. Each time the diaphragm moves out, it squeezes the air molecules immediately in front of it closer together, forming a region of high pressure. These molecules press on their neighbours, moving them closer together in their turn, and so a pulse of close-together molecules (a compression) moves through the medium, followed by a low-pressure area (rarefaction), which is produced as the diaphragm moves inwards.

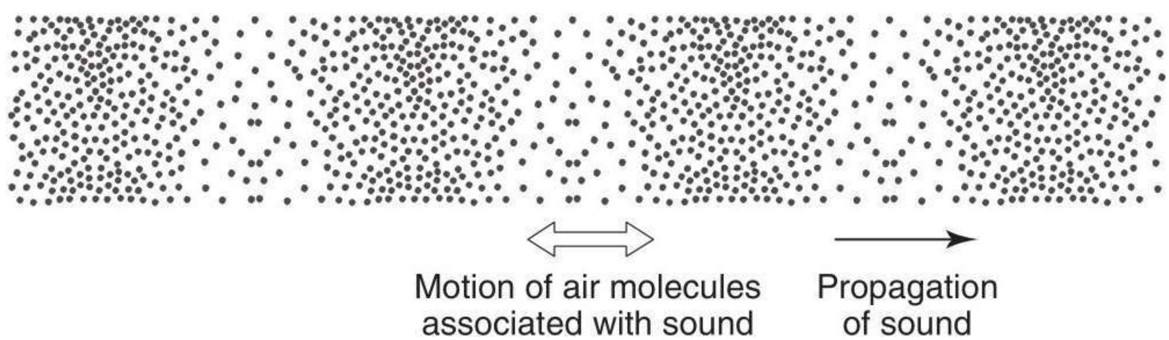
The diaphragm then moves out again, making a second pulse. How often the diaphragm moves in and out during 1 second gives the frequency of the sound wave (in hertz, abbreviated Hz). The simplest sound wave is a pure tone as made, for example, by a tuning fork; a snapshot of the variation in air pressure with distance from such a fork would be a sine wave, as shown in Figure 1.

The distance between adjacent peaks (or troughs) of a sound wave gives the wavelength (λ). The sound will travel through the air at a velocity v, which will be around 340

metres per second at room temperature. The frequency (f) is given by the equation $f=v/\lambda$. A plot of the variation of pressure over time at a single point in space is also a sine wave, so we could label the above x-axis of Figure 1 as ‘time’ if we wished.



1. Sound wave pressure plot.



2. Molecular view of sound wave.

Images like Figure 1 are so commonplace that it is easy to imagine they provide some kind of a picture of a sound wave, and many books use them in this way. In fact, however, there is no up and down (transverse) motion in a sound wave as there is in, say, an ocean wave—the only motion is of molecules shuttling alternately away from and towards the source, like the balls of a Newton’s cradle. Such waves are

referred to as longitudinal, and if we could see air molecules they would look something like Figure 2.

If a continuous sound originates from a point then it spreads in all directions, as an expanding sphere. If the detection area is small (like a microphone diaphragm or eardrum) and is several metres from the source, the curvature of the sound sphere is negligible, in which case the sound arrives in the form of plane waves. Even if the source of the sound has a particular direction to it (like most loudspeakers), the sound will still spread spherically, so long as the diaphragm is larger than the wavelength of the sound. For shorter wavelengths, the sound retains its original direction to some extent, and at suficiently high frequencies will form a beam (see Chapter 6).

Carrying sound

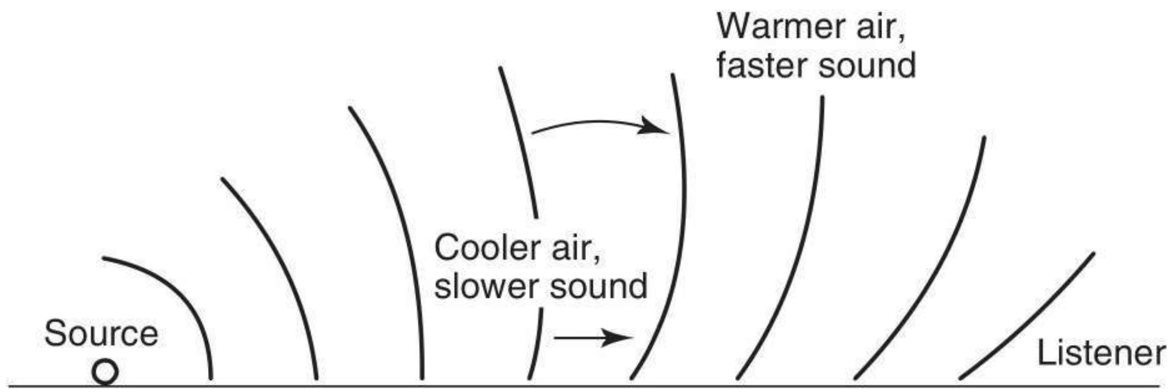
The velocity of sound depends only on the elasticity and density of the medium (Chapter 1). In air, sound velocity increases with increasing humidity and decreases with increasing temperature, but only because of the changes in density that these factors cause. Some examples are given in Table 1.

Material	Velocity of sound (metres per second)	Velocity of sound compared to that in air at 20°C, 80% relative humidity
Hydrogen at 0°C	1,286	3.73
Air at 0°C, 0% relative humidity	331.45	0.962
Air at 0°C, 80% relative humidity	331.70	0.963
Air at 20°C, 80% relative humidity	344.37	1
Water at 25°C	1,493	4.33
Seawater at 25°C	1,533	4.45
Diamond	12,000	34.85
Granite	5,950	17.28
Iron	5,130	14.90
Vulcanized rubber	55	0.16

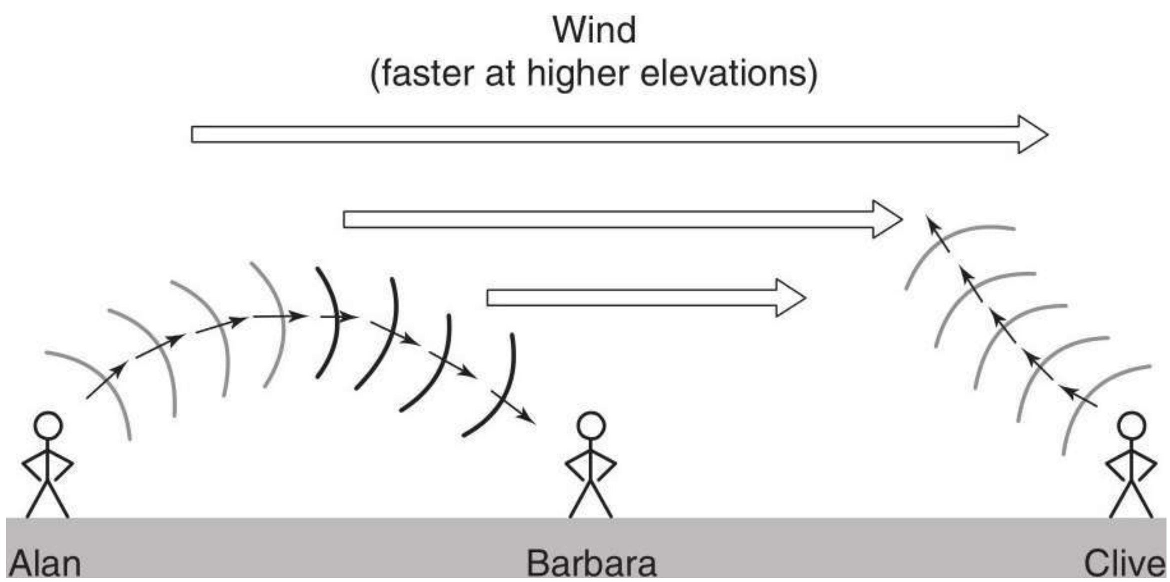
Table 1. Sound velocities in difierent media and conditions

Because the velocity of sound in air increases with temperature, on days when the air many metres up is hotter than that near the ground, sound waves travel faster the higher they are. The effect of this velocity increase is to bend (refract) them downwards from the warmer air, returning to Earth some distance away, as shown in Figure 3. Sometimes sounds can be heard more clearly at a great distance than at a short one due to this effect. Refraction also explains why it is hard to hear against the wind: the wind near the ground slows the sound waves slightly (compared to the ground) but the air a few metres up is faster, so the sound waves are

slowed a little more there. The sound refracts from the low-velocity region to the higher-velocity one, and hence curves up away from the ground and from your ears (see Figure 4).



3. Sound propagation when air is cooler near the ground than high above it.



4. Barbara can hear Alan, but not Clive.

No matter what one does with a diaphragm, one cannot push sound through the air that surrounds it any faster. More

rapid waggling generates pressure pulses that are closer together, which then arrive at a point—your eardrum, say—more frequently. That is to say, the sound frequency would rise. If one tries pushing the air harder by moving the diaphragm further in and out, then the amount of compression (and rarefaction) in the pulses increases, leading to a higher sound pressure (heard as a louder sound). If one forces the diaphragm to move faster than the velocity of sound in the medium, a pulse has no time to move away from the diaphragm before the next pulse forms. Hence, they pile up into a single, extreme high-pressure pulse known as a shock wave (the cause of sonic booms and whip cracks).

Moving a diaphragm more rapidly is not the only way to increase the frequency of a sound: if the loudspeaker (or other source) is rapidly approaching you or you are rapidly approaching it, the pressure pulses arrive at your ear more frequently, because each starts closer to you than the one before, so the frequency of the sound rises. Once the source has passed by, the pulses hit your ear with longer intervals in between, because each has a little farther to travel than the one before. So the frequency falls. This is the well-known Doppler effect, which happens whenever one is passed by a speeding motorbike or the siren of the police car behind it (Box 2).

Like light, sound reflects, and as with reflection from a mirror, an image of the sound source is formed if a surface is smooth and hard, so if you are somewhere between source

and surface, you will hear approximately the same sound from each side (but quieter from the reflector side). ‘Smooth’ is a relative term though, meaning ‘with bumps smaller than the wavelength’. Since sound waves are about one million times longer than light (comparing3 kHz to yellow), even quite rough surfaces like concrete make good acoustic mirrors. Concave acoustic mirrors focus the sounds that they reflect: in World War I just such concave concrete sound mirrors were built along the south coast of England to focus the sound of approaching planes into the ears of listening solders. When sound echoes between two or more curved reflectors, the result can be a whispering gallery, like the one in London’s Saint Paul’s Cathedral.

Box 2

Doppler effect: for an approaching sound source, the frequency heard by the observer is $f_{obs}=\left(\frac{v+u}{v}\right)f$; velocity of sound v, speed of approach u, frequency of source f. For a receding sound source, the ‘+’ becomes a ‘−’.

Sound will reflect from the interface between any two media, whether air and concrete, water and air, or different rock layers in the Earth. How much of the sound is reflected depends on the difference in the acoustic impedances of the two media, and the impedance in turn depends on the density of the medium and the velocity of sound in it. Acoustic impedance (Box 3) is similar to electrical resistance in that it measures the dificulty with which sound can travel through a medium. It is key to many of the effects and applications of sound. For instance, a soft rubbery surface will absorb sound and convert it to heat, since soft rubber has an

extremely high acoustic impedance. Stealth coatings on submarines are based on this fact, but unfortunately the softness of rubber is very temperature-dependent, so, when Cold War submarines were redeployed from the North Atlantic to the Gulf from the late 1980s, the higher water temperatures robbed them of their stealth and set off a flurry of research and resurfacing.

Sound can be focussed by passing it through an acoustic lens, often made of acrylic plastic. Lenses work because a wave is refracted when it passes from one medium to another, so long as it strikes the interface between the media at an angle. The angle through which the wave is refracted depends on the ratio of its velocities in the two media (Snell's law, Box 4).

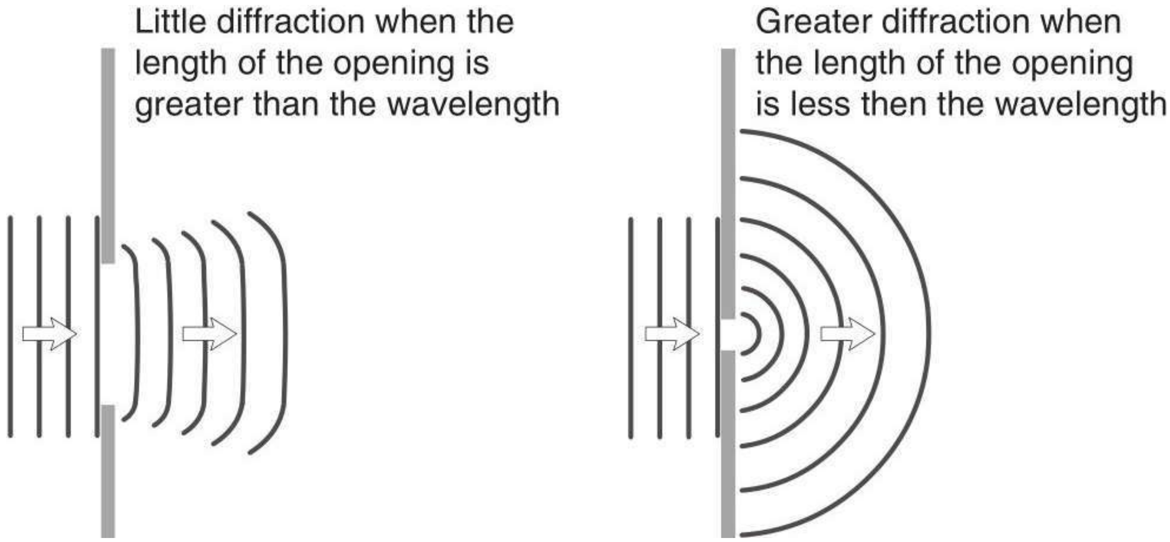
One effect which is usually far more noticeable for sound than for light is its ability to bend round corners and over walls, and to spread out after passing through an opening, a phenomenon known as diffraction or scattering (Figure 5).

Box 3

Characteristic acoustic impedance of a medium: $Z_0=\rho_0 v_0$; density ρ_0, velocity of sound v_0. The unit is the Rayl, and the $_0$s indicate that these are the values of the medium when it is 'unperturbed'—that is, when no sound is present in it.

Box 4

Snell's law: $\sin\theta_1/\sin\theta_2=v_1/v_2$; θ (theta) is the angle between the direction of the sound and a line at right angles to the interface, v is sound velocity.



5. Difiraction.

The longer the wavelength, the greater the bending. So if a high wall is introduced between a sound source—a band, say—and a listener, the bass sounds diffract from its top back down to earth but the high-pitched ones are lost (Figure 6). This muffling effect is a useful clue that helps us gauge the distances of familiar sound sources outdoors.

When light falls on a series of parallel lines, stripes, or ridges a single wavelength apart (or thereabouts), it is diffracted, and since shorter wavelengths are diffracted through larger angles, such difiraction gratings split white light into its component colours—the back of a CD makes rainbows from sunbeams in just this way. Since a pure tone is a regular series of 'stripes' of increased pressure, it can also act as a diffraction grating, scattering light with a wavelength around that of the distance between the stripes (that distance being half the wavelength of the sound).

Usually the medium here is a crystalline solid, such as fused quartz. This acousto-optic effect, where sound waves scatter light, is used both underwater and in air as a non-perturbing measurement and imaging tool (see Figure 7).

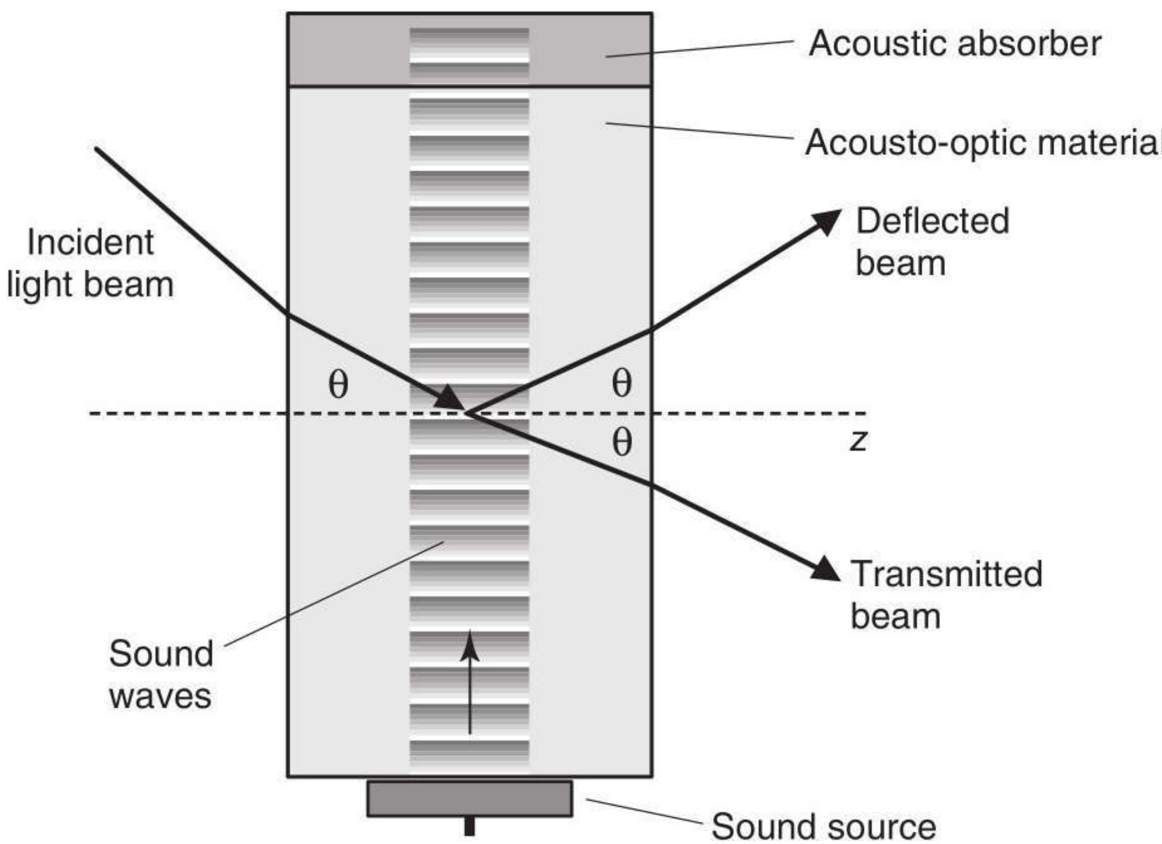


6. Difiraction of difierent wavelengths.

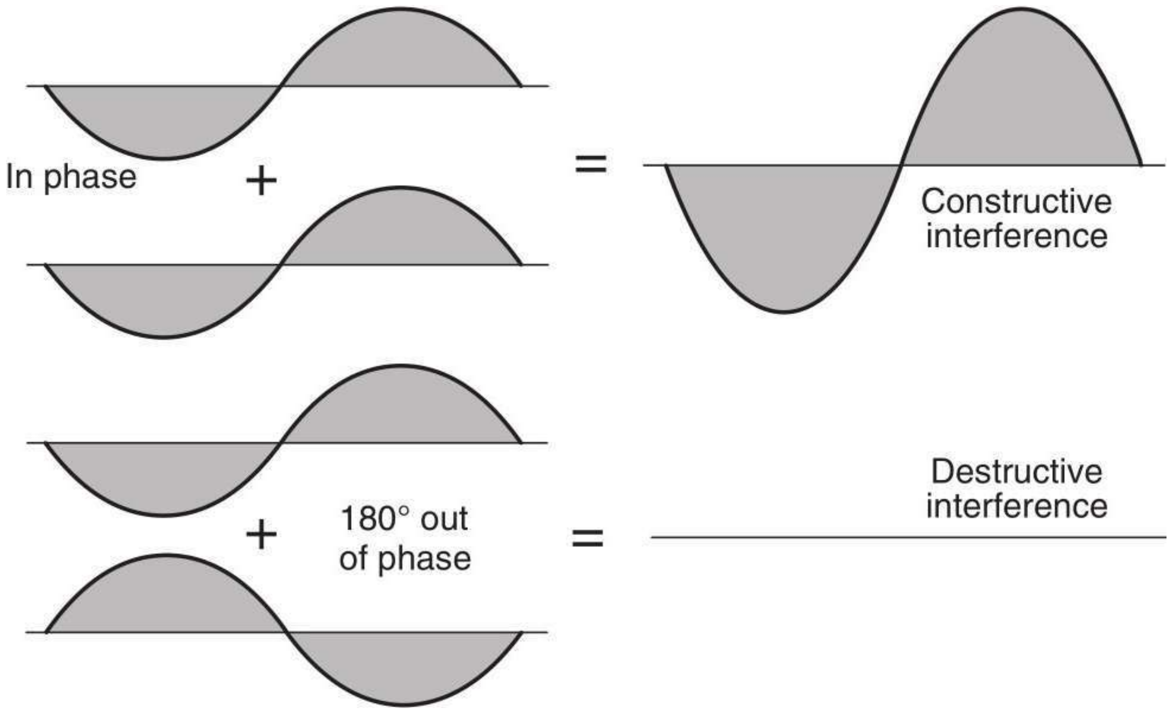
When sounds from multiple sources meet, mix, and mingle, the result is a three-dimensional pattern of loud and quiet areas called an interference pattern. The quiet areas form at the points where rarefactions from one source meet compressions from another (destructive interference), and the loud ones arise when rarefactions meet rarefactions, or

compressions meet compressions (constructive interference) (Figure 8).

Interference is important in stereo sound production and in noise cancellation, and it introduces one more parameter that characterizes a sound wave: its phase, that is, how high or low its pressure is at a particular point in space and time. Phase really only matters when sound waves interact: in the above example,



7. The acousto-optic efiect.



8. Constructive and destructive interference.

pairs of sound waves whose compressions coincide with each other (and hence make a loud area) are in phase, while those in which they do not coincide are out of phase. When waves are maximally out of phase, they are said to be in antiphase. Our hearing systems cannot detect phase.

The power of sound

There are several ways in which the amount of sound can be defined and measured, and each way is appropriate for different applications. If hearing or music is the context, sound pressure is the obvious choice since it is the parameter which relates most directly (though none too simply: read on!) to the impression of loudness. But in

discussing the eficiency of a sound source one may wish to know how much energy is flowing from it in a second—the sound power. To describe the effects of a particular sound field on an object, the parameter of interest is the sound intensity, which is the amount of sound energy striking 1 square metre of that object each second. Volume is an ill-defined measure used to label audio equipment, but intended to mimic loudness.

Audible-frequency sound waves lose very little energy through absorption by the air through which they pass (around 0.25 dB/6 per cent per 100 metres, though varying greatly with weather conditions). The main reason sounds die with distance is that they are free to spread out in many directions, so their energies spread progressively more and more thinly to occupy larger and larger volumes. If a sound source is suspended in free air so that its sound can spread in every direction (spherical spreading), then the sound pressure is inversely proportional to the distance of the listener from the source: that is, if the distance from source to measurement point doubles, the sound pressure halves.

The intensity of the sound falls more rapidly than this: it is inversely proportional to the square of the distance. So, if the distance from source to measurement point doubles, the sound intensity falls to one-quarter (1/22). If the distance is multiplied by 10, the intensity falls to one-hundredth (1/102). If the sound source is on the ground, the

waves spread hemispherically (Box 5), and sound pressure and intensity fall at half the above rates: in other words the intensity falls to one-half if the distance doubles—roughly,

Box 5

Spherical spreading: I=P/4pr2; intensity I, power P, distance from the source r.

unless the ground is a perfect reflector (a marble floor is pretty near), intensity will fall faster than this, due to loss of energy due to absorption. The sound power depends only on the source, so it is the same at any distance.

Pure tones are never found in nature, but are perhaps most closely approximated by the songs of birds. The waveforms of real sounds look very different: the pressure-plots of different sounds of similar fundamental frequency are shown in Figure 9.

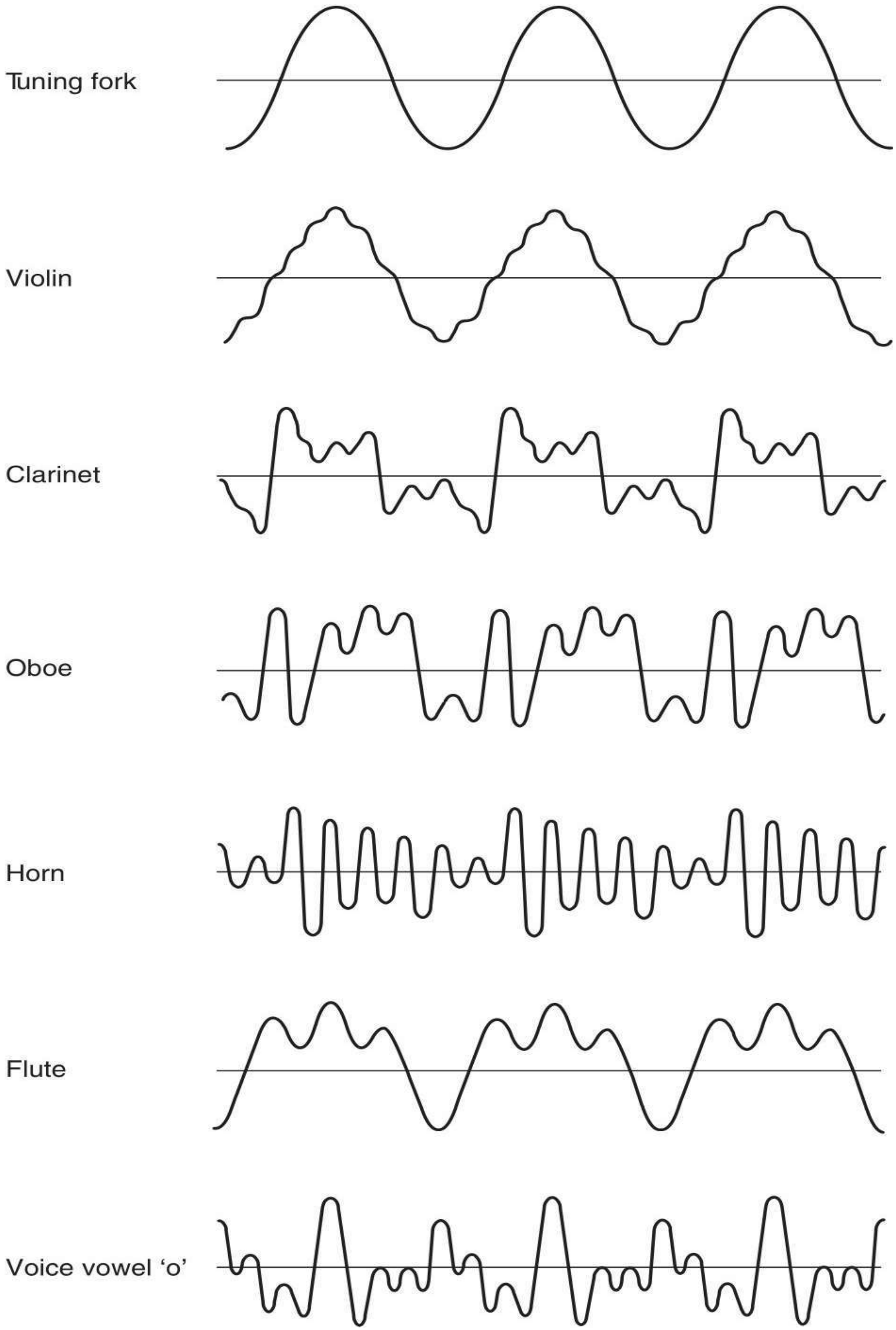
The difficult decibel

Sound was one of the first forms of energy to be understood: as far back as 300 bce it was known to be some kind of pattern of physical changes that could travel through air and water, but it was long after that that the most obvious characteristic of sound—its loudness—was quantified in any way. What did turn up took over 2,000 years to arrive, and was not very satisfactory when it did.

By far the most widely used way to quantify the amount of sound is the decibel (dB, Box 6): if two signals differ in

sound pressure by 1 dB, the ratio of those pressures is about 1.2:1. (Handily, this happens to be about the smallest difference we can hear under ideal conditions.) A 10 dB difference corresponds to a ratio of about 3:1, and a 100 dB difference to a ratio of 100,000:1.

A decibel is one-tenth of a bel, a name made by combining three letters commonly used in transmission theory ((β, ε, and l) with a tip of the hat to Alexander Graham Bell. Decibels aren't units—they are ratios, so they can describe how much more powerful one thing is than another: you could, if you wished, use them to compare the outputs of a pair of heaters. But that would not tell you anything about how hot either of them actually are.


Tuning fork
Violin
Clarinet
Oboe
Horn
Flute
Voice vowel ‘o’

9. Wave shapes.

Box 6

Decibels: the sound power difference in decibels between sounds of powers P_1 and P_0 is 10 $\log_{10}(P_1/P_0)$. Power is proportional to pressure squared, so the sound pressure difference between sound pressures p_1 and p_2 is 10 $\log_{10}(p_1^2/p_2^2)$, which is20 $\log_{10}(p_1/p_2)$.

To describe the sound of a device in decibels, it is vital to know what you are comparing it with. For airborne sound, the comparison is with a sound that is just hearable (corresponding to a pressure of twenty micropascals). When the amount of sound is given in terms of such a reference level, the word 'level' is appended: hence sound pressure level (SPL), for example. So, a sound of 0 dB is 'one times louder than' (i.e. equally loud as) a sound you can just about hear, 1 dB is about 1:12 times as loud, 2 dB is 1:26 times, and so on. Are all acousticians happy with this solution? No, they are not. Ultrasound engineers don't care how much 'louder than you can just about hear' their ultrasound is, because no one can hear it in the first place. It's power they like, and it's watts they measure it in. Meanwhile, underwater acousticians rightly ask: 'The threshold of hearing? What does that mean when your ears are full of water and you've got a rubber headpiece on? Or if you're a whale?' So they base their decibels on a reference pressure of one micropascal, because that's nice and easy to remember. So now we have two kinds of decibel, one for use in water and one for air, which will give different values for

the same sound. Not too much of a problem so long as everyone always remembers to say what the reference level of the decibel they are using is. Sadly however, they don’t.

There is another problem. Few of us care how much sound an object produces—what we want to know is how loud it will sound. And that depends on how far away the thing is. This may seem obvious, but it means that we can’t ever say that the SPL of a car horn is 90 dB, only that it has that value at some stated distance. Often, even those handy decibel charts so popular in textbooks get this bit wrong, and claim the SPL of a pneumatic drill is 100 dB, when they really mean ‘100 dB if measured at a distance of 10 (or however many) metres’. It’s not dificult to see where this laziness creeps in—the charts usually also have examples like ‘a quiet ofice’, and we understand that the chart maker is referring to a quiet ofice that you are working in, not a quiet ofice down the corridor or in another town.

There is a third problem: a source of sound might make sound waves at any one frequency, at a handful or at a wide variety of frequencies. Let’s assume for the moment that the source of the sound is a loudspeaker so eficient that it always converts all the electrical energy fed into it into sound. And let’s imagine it has a frequency control but no volume knob. If we measured the total sound energy flowing from that loudspeaker each second (that is, the power) while changing its frequency, that power would of course remain constant. Similarly, the SPL at a particular distance from

the loudspeaker would stay the same—as a microphone would show (assuming it were equally sensitive at all frequencies).

However, this is nothing like what your ears would tell you. If the loudspeaker was just audible at 20 Hz, it would increase in loudness as the frequency rose, until at around 4 kHz it would sound (very roughly) 200 times louder. At higher frequencies still, it would get quieter again, finally fading into inaudibility at somewhere between 8 kHz and 20 kHz, depending how old you are and what you' ve been doing to your ears for the past few decades.

In practice, acousticians weight the response of the circuit of which the microphone forms a part, so that the system behaves like the ear—being most sensitive to frequencies at around 4 kHz. A frequency-weighted microphone is the heart of a sound level meter (SLM). There is actually a wide choice of different weightings, including some for dogs, but the most popular by far is A-weighting, which approximates the response of a human ear at moderate sound levels. Hence, the decibels that matter to us are usually A-weighted, written dBA, the full name for which is 'A-level weighted sound pressure level in decibels' .

SLMs are equipped with a choice of time integration factors. This matters because if a sound lasts less than about 0.1 second it sounds quieter, since the hearing system adds up the energy for about this period.

To add to the complexity, the loudness of a sound also depends on the nature of its source. For instance, people so dislike the sound of planes that, on average, they consider them to be as annoying as anonymous sounds that are about 5 dB louder. Conversely, people are rather fond of train noise, to the extent that they only find it as annoying as anonymous sounds which are about 5 dB quieter. These reactions are so well established that many planning applications which involve aircraft or railway noise adjust their figures by 5 dB (the corrections are known as the aircraft malus and the railway bonus). This means that no meter can actually measure what architects, home owners, noise campaigners, noisy machine buyers, and acousticians really need to know: how loud a sound is.

Considering all of this, there is little point in measuring SPLs with great accuracy: most SLMs are accurate to +/- 1.4 dB at10 kHz (called Class 2 meters). Even in laboratory work, +/- 1.1 dB at 10 kHz is almost always enough (as provided by a Class 1 SLM). Far more important than accuracy is adherence to standard measurement procedures, including the frequent calibration of SLMs by comparison with standard measurement microphones.

Despite the complexities of loudness and its variation according to the source and the user, extensive surveys of the reactions of large numbers of individuals to carefully chosen sounds have determined roughly how loudness relates to SPL, and units have been defined on this basis, in particular

the phon. Phons are defined as having the same values as the SPLs of 1 kHz tones, so a 1 kHz tone with an SPL of 10 dB has a loudness level of ten phons. But a 50 Hz tone with that same loudness level of ten phons has an SPL of 73 dB, because our ears are so much less sensitive at 50 Hz than 1 kHz that a 50 Hz tone needs to be 63 dB higher than a 1 kHz tone to sound as loud.

Loudness is just one of a large set of psychoacoustic measures, also known as sound quality parameters (‘quality’ being used in the sense of ‘character’ rather than ‘goodness’). Loudness is by far the most commonly used and best developed; the others include sharpness (in acums), roughness (aspers), fluctuation (vacils), and dieselness (which has no units: different automobiles are simply ranked subjectively according to how ‘dieselly’ people think they sound). As the last-named suggests, these measures were developed primarily by the automotive industry in its attempts to make door-clunks, engine sounds, and even indicator noises sound appropriately powerful, masculine, reliable, and so on. In principle it would be very useful if domestic products and other noise sources could be characterized by such parameters.

The topic of sound qualities is part of the discipline of psychoacoustics, the study of the psychological effects of sound, which itself can be considered as an element of what is now known as sound studies. Sound studies deal with how sounds of all kinds have been made and consumed throughout

history and in different cultures. Work on such topics has been carried out since the 1940s, and has increased greatly since the early 1990s.

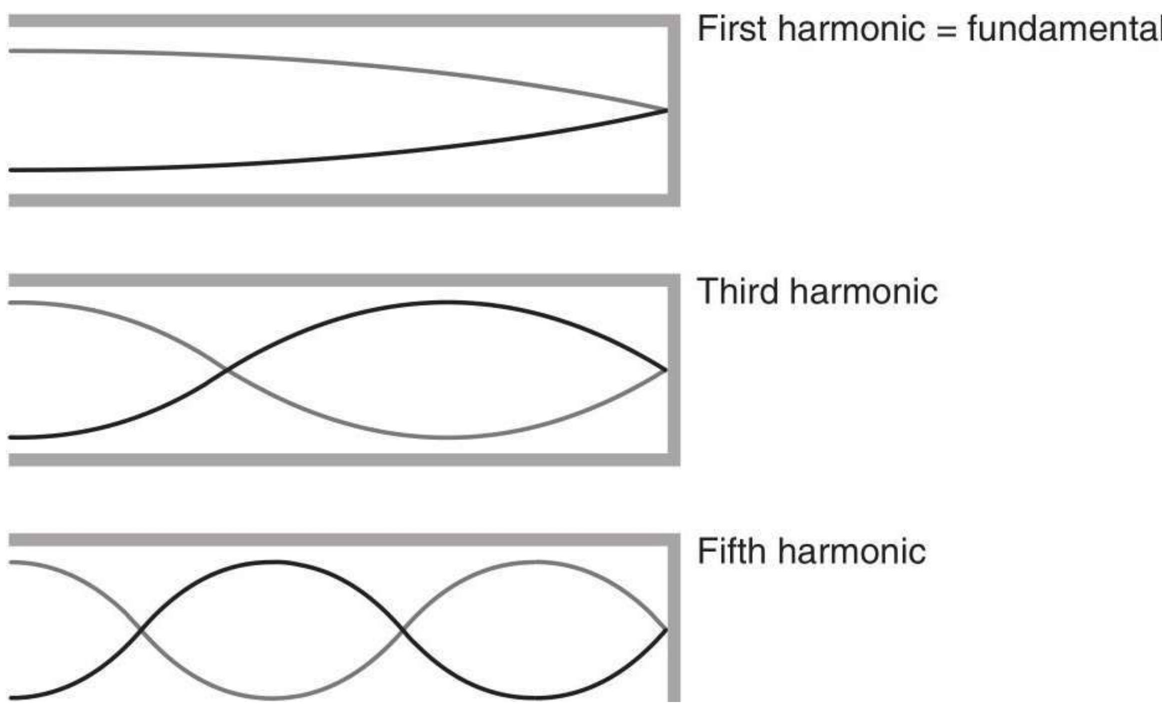
Standing waves

A recurring aim in the history of acoustics is to make sound visible. In the 1780s, Ernst Chladni studied the ways in which metal plates vibrate when they are made to 'sing' by being stroked with a violin bow. Fine powder sprinkled on the plates is deflected from areas where vibration is strong, and collects in those that are still. The powder-free, strongly vibrating points correspond to antinodes (like the peaks or troughs in Figure 1), and the stationary, powdery areas are the nodes, the points where there is no pressure change (where the line crosses the axis in Figure 1).

It was possible for Chladni to 'see' sound waves in this way only because they did not progress through space: they were stationary, or 'standing' waves. For a standing wave, Figure 1 represents only how the pressure of the wave changes with location, and not the way the pressure at a particular spot changes over time (such a time-based diagram for any point in a standing wave would be a horizontal line).

The principle is clearer when considering the standing sound waves that are formed when one blows across the open end of a12 cm tube which is closed at the other end. In all such waves, the air at the closed end of the tube cannot move, due to friction with the end wall (so this point is a

node). The simplest wave of this kind is one in which the motion of the air molecules increases with distance from this end, and reaches a maximum (an antinode) at the open end. In this wave, one-quarter of a wavelength is inside the tube, so it has a wavelength of 4 × 12 = 48 cm. If one blows hard enough, a whole range of other standing waves will form, each with a node at one end of the tube and an antinode at the other, as shown in Figure 10. The lengths of these other waves are simple multiples of the first, and such waves are known as harmonics.



10. Standing waves in a pipe open at one end.

Just as in such a pipe, so in any other fluid-filled cavity, or any rigid object, there are certain wavelengths of sound which are particularly easy to excite. These are called

resonance modes (or simply resonances), and the main ones can be predicted since they depend only on dimensions. For instance, if a 12 cm rod is fixed at its ends and struck sharply, it will generate 24 cm sound waves, together with waves of lengths 12 cm, 8 cm, 6 cm, 4 cm, and all others which include nodes 24 cm apart—again, a set of harmonics.

A 12 cm long box of air or water will produce all these waves too—in this case, what stops the 'ends' of the fluid moving is that they are adjacent to the box walls, where friction prevents free motion. The box will also produce families of waves corresponding to its height, width, and diagonals.

Resonances can be a major problem in room acoustics, but are the basis of most musical instruments. With an instrument that includes tubes open at one end (like some organ pipes), the open end is an antinode, and hence the fundamental frequency has twice the wavelength of that of a closed tube of the same length. (Actually, the antinode forms just beyond the pipe' s end, requiring an end correction to be made, see Box 7.)

Box 7

End correction: $\frac{\lambda}{4}$=l+c (closed pipe), $\frac{\lambda}{2}$=l+2c (open pipe). End correction c,4 about 0.6 of the pipe ra2dius.

Usually, the lowest resonant frequency is the most powerful; however, if a great deal of energy is supplied to an instrument, it may resonate an octave—or even two—

higher. A flute, for example, will do this if it is blown hard enough (‘overblown’).

Resonances are all around us—tap a plate, glass, or fork and it will ring, provided only that it is not damped by being held too tightly (tuning forks still resonate if held tightly because they have two identical prongs that move in opposite directions, cancelling each other out at the handle so there is no resultant motion there). This is a handy way of finding whether crockery is cracked: if all is well, each successive millimetre of the plate will move immediately after its neighbouring millimetre, allowing waves to pass, much like a Mexican wave: the plate is, literally, sound. But if even a very line crack separates adjacent areas, dragging and friction damp the resonance, giving rise to an unhealthy ‘clink’ .

If the force supplied to an object is a sound at the resonant frequency, the coupling will be highly eficient, hence guitar strings that sound in sympathy with those struck across the room, or bits of television sets that buzz annoyingly along with dramatic programme sounds.

An effect which is of importance in several areas of acoustics is Helmholtz resonance, familiar to anyone who has blown across the top of a bottle to make it sing. Any hollow object or cavity with an open neck will act as a Helmholtz resonator (Box 8). If a stream of air is blown across the opening, some will enter the neck, increasing the pressure in

the cavity a little. This overpressure pushes the air out again—and, just like a pendulum,

this air ‘overshoots’ a bit, leaving a slight underpressure, which sucks more air in, and so on. This regular cycling constitutes a sound wave at a resonant frequency. If a sound wave at this frequency is supplied to the resonator, it will sound very strongly.

Box 8

Helmholtz resonator frequency: $f \approx \frac{v}{2\pi}\sqrt{\frac{S}{lV}}$S; sound velocity v, neck area S, neck length (including end correction) l, cavity volume V.

Charting sounds

Standing waves are a small subset of sound waves: mostly, the regions of high and low pressure in a wave move through space (such waves are called progressive or travelling waves). If one wishes to ‘see’ a travelling wave, one must therefore chart air pressure changes through time. One of the first to attempt this was Alexander Graham Bell, who in 1874 procured an ear from a corpse, impregnated it with oil to keep it flexible, and attached a thin straw to its drum. The other end of the straw was allowed to trace a line on a strip of soot-covered glass which was moved along as the ear was shouted at. This wobbly line was the first recording of a sound wave and the device was called an ear phonautograph. To the relief of those who had to construct them, later versions dispensed with dead ears in favour of metal diaphragms.

Phonautographs were no good for making actual measurements of sound waves, however; these were eventually provided by the cathode ray oscilloscope (CRO), developed in the 1930s. CROs can be set with different time-bases so that a high-frequency sound can be spread out across the screen or a low frequency one compressed, so that their wave shapes can be seen. From this, their wavelengths can be read off and their frequencies determined.



11. Spectrogram.

Today, computerized versions of CROs are widely used. However, a two-dimensional plot can still only display some features of sound. Most sound waves vary rapidly both in frequency content and in pressure, which can only be properly displayed together on a three-dimensional display, called a spectrogram, which cannot be produced without a computer. In a spectrogram, height up the screen usually represents

frequency, and brightness or colour represents sound pressure (or intensity). In other cases, a representation of a three-dimensional shape may be made on a screen, the results often resembling mountain ranges (Figure 11).

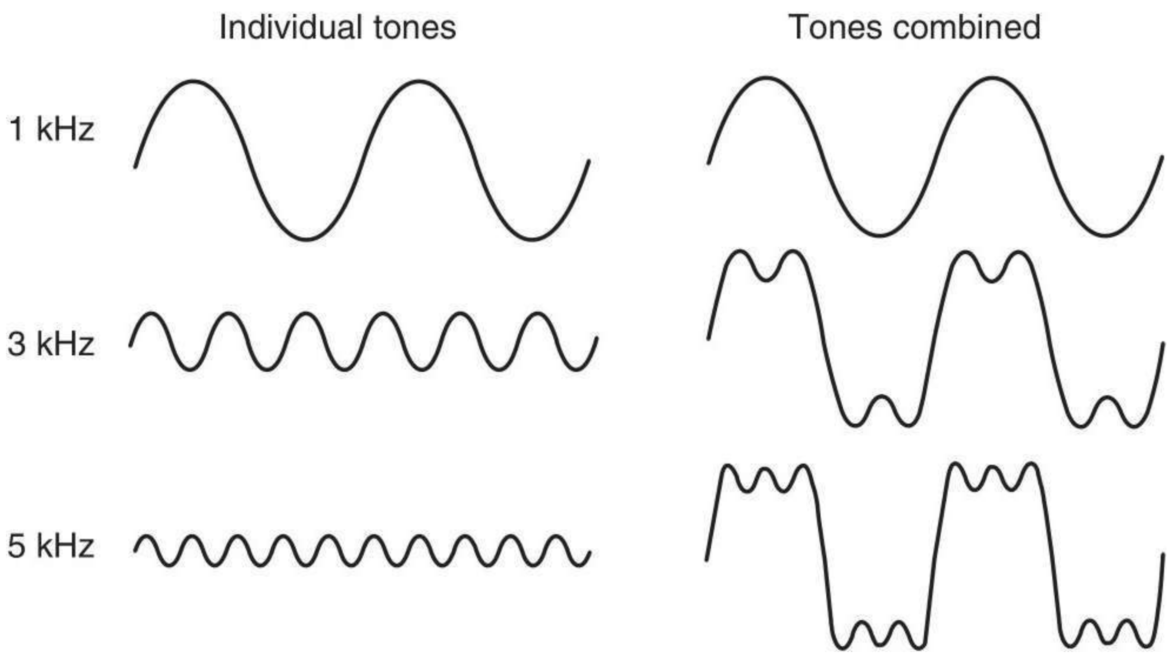
Unweaving sounds

Being able to see a sound allows one to find out a lot about it qualitatively, and rough measurements can also be made of the screen outputs, but often good quantitative information about sound is needed (perhaps to eliminate noise or improve the design of a musical instrument). For this, a mathematical analysis is required, and the most widely used and fundamental is based on work conducted by Joseph Fourier in the 1800s.

Fourier realized that any periodic function (that is, one which repeats at a steady rate) can be constructed by adding together a series (now called a Fourier series) of sine waves—and he worked out a method to determine what the members (terms) of that series are. (Mathematically speaking a Fourier series is composed of a series of sines and cosines—but a cosine is simply a sine wave which starts at maximum, rather than at zero, so I'll just refer to sine waves here.) As Figure 12 shows, as few as three sine waves can roughly approximate a square wave.

To make the sides of the latter more vertical, higher-frequency tones must be added. A square wave sounds like a click, and Fourier analysis shows therefore that a sudden

(that is, rapidly increasing in level) click will include some very high-frequency components.



12. Summing sine waves to approximate a square wave.

Fourier’s original work was only applicable to periodic waves, but a development of it known as the Fourier transform can handle non-periodic ones. A highly eficient mathematical method of calculating the component sine waves of a signal is known as a Fast Fourier Transform (FFT). When adding waves like this, their phase must be taken into account. During a single wavelength the sound pressure of a wave rises from zero (that is, equal to the ambient air pressure) to maximum, then falls down to minimum, and then rises to zero again. This is similar to the vertical motion of a dot painted on the edge of a rotating wheel, so phase can be described in terms of angles: starting at 0° the wave rises to a maximum

at a phase of 90°, falls to zero at 180°, down to a minimum at 270°, and back to zero at 360° (which is the same as 0°).

All real sounds change over time, so the conversion into sine waves must be repeated frequently. Such time-varying frequency analyses of sounds have many applications: for instance, some of the parameters of the sound waves which compose an individual's voice are unique to that person. Hence, such parameters can be used as acoustic 'fingerprints' (called voice prints), and automatically recognized by a machine.

Conversely, since every word has a unique sound (except for homophones like 'sew' and 'so'), it is in principle possible for a machine to recognize them automatically, whoever speaks them. While the same word will be said differently by different speakers there are certain elements which vary only slightly, or predictably—hence our ability to recognize a word irrespective (within limits) of its speaker.

Automatic speech recognition is a long way from perfection however, and the main problem lies in deciding where one word ends and the next begins. To see how tricky this is, try listening to yourself saying 'bread and butter'. You will probably actually hear something like 'brembudder' with no silences at all (and saying the phrase 'properly' neither feels nor sounds natural). The reason

that humans can identify words so readily is that the sound patterns we hear are only one piece of evidence as to what is being said—as Chapter 4 will explain.

Sounds from nowhere

Since any sound can be analysed into sine waves, it follows that any sound can be synthesized from them: synthesizers that generate speech from sounds have been available for many years, and work far better than recognizers. In practice though, it is often far easier to generate speech by adding together fragments of pre-recorded or pre-generated sounds—a technique known as voice coding.

Today's electronic systems can synthesize practically any sound at all, whether it occurs in nature or not—like the weird Shepard tone, which is produced by combining tones which fall in pitch but then fade out, while other, higher tones, fade in and themselves begin to fall. The impression is of a sound which continually falls—and yet gets no lower.

Usually, however, one does not want new sounds but improved versions of existing ones—a musical performance shorn of noise, for example. The selection of pre-recorded elements is also commonly used for non-speech sounds: one of the most celebrated electronic gadgets for the budding pop music producer in the 1960s was the Mellotron (1963)—a machine loaded with a library of short pieces of sound recorded on to magnetic tape, any of which could be quickly selected and played at a chosen frequency and volume.

Selecting sounds: filters

The commonest and easiest way to modify sound is through filtering: the removal or reduction of selected frequency ranges, accomplished either by electronic circuits or by software. High-pass filters cut out low frequencies, low-pass ones deal with high frequencies and band-pass filters banish both. A once familiar kind of variable filter is the graphic equalizer, a series of about seven slider bars on a hi-fi’s amplifier, which allows selected preset frequency ranges to be suppressed. The simpler ‘tone’ control similarly quietens either high (‘treble’) or low (‘bass’) frequencies.

A vast range of other facilities is also to be found in the computerized toolbox of the sound artist or engineer: such software can add reverberation or echo to a recording, create an artificial soundscape, or apply such real-time changes as shifting the frequencies of the sounds of a pop song recording before passing to a loudspeaker. This is the basis of a karaoke system, in which the notes of songs can be sharpened or flattened to match those that the user finds easiest to sing.

Chapter 3 Sounds in harmony

What makes a note?

The words 'tone' and 'note' refiect the subjective/objective nature of sound: a tone is a sound wave with a particular frequency, a note is its subjective impact, with a particular pitch.

In addition to pitch, a note also has duration, loudness, and timbre. Timbre is by far the most important in identifying instruments by their sound, and is also the main carrier of emotional content. It is actually an uneasy bedfellow with the others, being far more complicated. It is a bit like deffning a person by gender, height, weight, and ffngerprint: the ffrst three parameters can each be speciffed in terms of one value of a single unit, but the full description of an individual's ffngerprint would be highly complicated and multidimensional. But, like a ffngerprint, timbre is the only thing unique to a particular instrument or person, so it's the only thing that the brain can use as an identiffer.

A problem for the brain which the ffngerprint detective need not worry about is that timbre is a dynamic quality; it changes over time. The timbres of a cymbal, piano note, or drum all change radically from start to ffnish (this change is known as fiux). The timbre of a particular instrument or speaker also changes with pitch and loudness: a cello's high

notes have a diflerent quality to its low ones, a shouted word is not simply louder than a spoken one, and a man's voice at the top of his range sounds much thinner than when he is singing low notes. Fortunately, we can rely on our highly evolved ability to focus on the very ffrst sounds made by a source to untangle all this.

Prehistoric agendas

Our modern appreciation of music and our dislike of noise probably stem from the evolutionary pressures that moulded the hearing systems of our distant ancestors. Their ffrst priority on hearing a sound would have been to identify it, and there was no time to ponder. The twang of a bowstring, the rumble of an avalanche, the thunder of approaching hooves or the warning hiss of a snake are only of benefft to the hearer if (s)he can react fast, and so the hearing system concentrates its analytical prowess on the earliest moments of a sound. This has the surprising consequence that the 'attack' sounds that an instrument makes in the ffrst fraction of a second, though generally discordant and nothing like the instrument's 'steady state' sounds, are the ones that enable us to identify what that instrument is. A recorded piece of music edited to remove all the attacks sounds very strange, and the instruments could be anything. This fact stymied for many years attempts to synthesize instruments convincingly.Often not just the instrument but the composition too can be identiffed within the ffrst second. There was once a fairly popular radio (and then TV)

quiz show called Name that Tune, in which contestants frequently succeeded in identifying tunes from just four or even three notes. It is very easy to outdo this if one is presented with a familiar recording: even a preliminary fumble or intake of breath is enough to identify it. Conversely, some musical sounds immediately inform the listener of the fiavour of the whole piece: a single chord may have a ‘Mexican’ sound, a few violin strokes might sound ‘folksy’, a bagpipe skirl ‘Scots’, or a couple of notes ‘fferce shark approaching’.

The responses to such brief cues are primarily emotional, as beffts their original role as hazard warnings: the roar of a lion or the click of a safety catch being released deliver an immediate visceral response, releasing adrenalin in preparation for fiight—or ffght. In fact, our refiex response to any sudden loud sound close at hand is to put distance between us and it: a feedback shriek from a loudspeaker will literally make us jump away. Even continuous sounds are intrinsically repellent if they are loud enough: hence the almost physical reluctance one feels to pass close by a pneumatic drill.

The fiipside to our focus on the very ffrst parts of a sound is disinterest concerning sounds that last for a while, leading to the seeming paradox that a continuous sound at constant volume will sound quieter after a few dozen seconds: in nature, such sounds are likely to be as harmless as the

breeze. But if that long-duration sound should suddenly stop, the silence is just as attention-getting as was its onset.

The act of listening to music has received considerable attention from a sound studies perspective. The musicologist and philosopher Peter Szendy, for example, argues that an essential part of listening to a piece of music is comparison with other works, performers, instruments—and other listeners too. His view is that, since the act of listening constantly 'appropriates' other things in this way, the essence of a piece of music can never be fully grasped.

Learning to sing

Musical sounds are far more important to many of us than those of other types—so much so that many of us can hardly restrain ourselves from bursting into song at the slightest provocation. Perhaps you have entertained yourself and, to a rather lesser extent, everyone within earshot, by singing 'Somewhere Over the Rainbow'. It's a fair bet that Judy Garland did it better: but how? Although your listeners may not enjoy the tune, they will surely recognize it because, even if a lot of the notes are wrong, they will at least go up and down in pitch in the right order, and vary in length in a familiar way.

Durations of notes (and of rests; gaps between notes) are indicated by their shapes, and deffned as fractions of a 'whole note', which is a semibreve, as shown in Table 2.

There is no standard duration for a note, though in some cases a composer will provide one on an individual score: ‘fi = 66’, for example, means ‘a crotchet lasts 1/66 of a minute’. Usually however, the only indication of the speed (technically, tempo) at which a piece should be played is a phrase (often in Italian) of variable helpfulness, such as andante (‘at a walking pace’) or allegro non ma troppo (‘fast but not too fast’). Perhaps because the concept of tempo is not very applicable to natural sounds, we are poor at judging it—amateur or professional alike, spotting a 4 per cent diflerence in tempo is about the best we can do.

Unless you are a trained singer, or a naturally good one, you are probably singing some notes slightly ‘fiat’ (that is, a bit lower than they should be) or ‘sharp’ (a bit higher). But what does ‘should’ mean here? It doesn’t mean ‘hitting the right notes’—it doesn’t matter whether the frequency of the sound you make while singing ‘Some . . .’ is 300 Hz or 333 Hz. But you do need to get the ratios of the frequencies correct. So, if you sing ‘Some . . .’ at 300 Hz then you must sing ‘. . . where’ at600 Hz—that is, one octave higher—to avoid people wincing.

Sign	Name	Relative length	Rest
𝅝	Semibreve	Whole note	𝄻
𝅗𝅥	Minim	Half note	𝄼
♩	Crotchet	Quarter note	𝄽
♪	Quaver	Eighth note	𝄾
𝅘𝅥𝅯	Semiquaver	Sixteenth note	𝄿

Table 2. Notes and rests

It follows from this that if you accompanied yourself on a piano without altering the pitches of your singing, you would almost certainly be out of tune with the instrument. As far as a pianist is concerned, ‘Some’ is an F. There are eight F keys on an eighty-eight-key piano, and any of them may be used here—if the fourth F is chosen, the piano will play a note of 349 Hz.

If you do ffnd yourself to be ‘automatically’ in tune with a piano, you probably have that rare thing: absolute (or ‘perfect’) pitch. Generating the same notes as your piano is all there is to having absolute pitch. If you were transported back two centuries, when pianos were not tuned to the same frequencies as they are now, you would of course not be in tune with them. And if you attempted to sing an

unaccompanied duet with a singer of the period who had perfect pitch, the result would not sound pleasant.

Your pitch is nevertheless better than your 19th-century friend's in a certain sense: if you travelled two centuries into the future, your perfect pitch would match everyone else's, but your friend's would not. That is because, in 1939, at a London meeting of the International Standards Organization, many countries agreed on particular frequencies for particular notes. Like many other international agreements, this one only happened because the situation had become intolerable some considerable time before.

Since instrument makers, musicians, and composers had developed largely independently in diflerent countries, each had adopted their pitches arbitrarily. In countries with several musical centres, pitches even varied from city to city. In 1780, a Berliner with perfect pitch would fall sadly fiat in Vienna; a trumpet made in one city would be out of tune in the other—and trumpets cannot be retuned. Hence, as national and international travel among musicians became common,

and successful instrument makers expanded their customer bases further affeld, the need for an international standard became rapidly more apparent—hence the meeting, even on the eve of war.

It follows that absolute pitch cannot be an inborn talent; it must be learned. And learned before the age of

six, usually by memorizing all the notes on a musical instrument. This is, on the face of it, an almost incredible feat of memory, and such memorization skills are alien to the adult mind. It seems that some children are predisposed to learn absolute pitch—but on the other hand there is a higher fraction of pitch-perfect persons in China and Vietnam than in Europe or the US, which may be due to the greater dependence in Eastern languages on the pitch at which words are pronounced.

For those of us who lack absolute pitch, it is the diflerences (intervals) between notes, not the notes themselves, that we memorize; hence a familiar tune is instantly recognized whether it is played on the double bass or the violin.

Why do we like what we do?

By far the most important interval in music is the octave, the interval in which one note has twice the frequency of the other. The primacy of the octave is demonstrated by the fact that, in the A to G notation system most of us use today, notes an octave apart are given the same letter. To distinguish diflerent versions of each note, numerical subscripts are used as required (so the fourth F on the piano is F4).Notes one octave apart are in harmony, and the impression we receive of a harmonious interval is that of auditory pleasantness: consonance. Other than a pair of

identical notes, the octave is the most harmonious and consonant interval of all.

Exactly why harmonious intervals sound consonant is not quite clear, though we do know that the brain responds to harmonic sounds with synchronous neural ffrings. That is, the neurons which respond to each component of such sounds tend to synchronize their ffring rates with each other. So, harmonious intervals are clearly important to us—but why do we ffnd them pleasant, and inharmonious (dissonant) intervals to be less so? Actually, though we don' t usually enjoy discords in isolation, the appetites of composers and their audiences for dissonance have grown steadily since the ancient Greeks—in fact, one could look at the history of Western music as the increasing adoption of dissonance (which is why listening to Stravinsky' s Rite of Spring or Mozart' s Dissonance quartet is not as disturbing as it once was). Nevertheless, if one ranks intervals by their perceived consonance, the result is pretty independent of geography or history. So there is there is something objective about consonance. In fact, two things:

1. A lack of 'roughness' . When the ear receives two tones close together in frequency (about 5 per cent diflerent, or two or three semitones), they are perceived as a single rough-sounding note.

2. Similarity to a harmonic series, which is a set of tones with frequencies such as 1 kHz, 2 kHz, 3 kHz, 4 kHz,

etc.

More notes

Music making cannot rely on octaves alone. For one thing, we can’t hear more than about ten of them. Also, producing such a range is a challenge: violins and guitars cover about four octaves, exceptional humans can manage ffve or six, pianos seven or eight. Furthermore, every tune would be nothing but the purest harmonies—which very soon become boring.

Once the octave had been deffned, Pythagoras and others set about adding notes to it, embarking on a long research project that was to last for well over a thousand years. Although Pythagoras actually used a monochord, it’s simplest to the job with several strings of variable lengths, all with the same tension and width and all made of the same material.

The next most harmonious/consonant pair of notes after the octave is that given by a pair of strings in which one string is 2/3 the length of the other. Three other simple ratios of string lengths yield other fairly nice-sounding note combinations, and the results are shown in Table 3.

Name of interval	Length of shorter string compared to first string	Frequency of note from shorter string compared to note from first	Order of consonance
Unison	1	1	1
Second	8/9	1 1/8	6
Third	4/5	1 1/4	5
Fifth	2/3	1 1/2	3
Fourth	3/5	1 2/3	4
Octave	1/2	2	2

Table 3. Harmonious intervals

If we were to build a harp with strings of lengths deffned in Table 3, but all of the same thickness, tension, and material, we would very easily be able to make simple, reasonably pleasant-sounding chords with it—in fact, any pair or larger group of such strings played together wouldn’t sound too bad.

This system is called the pentatonic (ffve-tone) scale. It was discovered independently by numerous ancient musicians in many parts of the world and is still very popular today (‘Amazing Grace’, ‘My Girl’, and ‘I Shot the Sherifl’ are written in it). The black keys on the piano are a series of pentatonic scales, so if you press them randomly it won’t sound as horrible as if you do the same with the white ones.

Even more notes

Although people like musical harmony, they also like a bit of discord, and the pentatonic scale was just too tuneful and safe for composers who wanted to challenge their audiences and push the boundaries. Also, ffve notes per octave is still not that many: but how to add more? Answering this question initiated centuries of careful experiment and academic debate, leading ffnally to a system called equal temperament. In this system, an octave is divided up into twelve notes, each a semitone, or one hundred cents, apart.

In keeping with the idea that what matters in music are ratios of notes, not notes themselves, semitones and cents are deffned in terms of ratios. Each successive semitone is about 6 per cent higher than the one before. This means that the higher the number of hertz corresponding to a semitone, the higher the note is.

A few composers have experimented with more than twelve notes per octave, and extra notes may appear briefiy, in moving gradually from one note to another to another (this is called a glissando, or slide). However, in all cultures, the number of 'proper' notes per octave is almost always twelve.

Having deffned twelve notes, people proceeded to use just eight of them per octave (hence the name). There are only seven difierent notes in an octave—the eighth is the same as the ffrst, just one octave higher (even professional

musicians who can immediately identify a note struggle to say which octave it is).

The reason why we use only eight of the twelve notes in an octave was resolved in 1956 by George Miller, a psychologist, who found experimentally that our short-term memory can store no more than about seven items—hence seven diflerent notes—at a time.

There is a semitone jump between adjacent notes in the full group of twelve, so any group of seven notes that we choose to make our scale will include some which are two semitones (one tone) apart. So, we might have this scale: A gap B C gap D gap E F gap G. (A scale is any sequence of notes in which each is higher than the preceding one.)

The fact that we have chosen a set of seven diflerent notes to deffne our scale does not mean that we may not use other notes in a piece of music. If we want to use those 'gap' notes, we mark them with symbols which mean 'a semitone higher than the next lowest note in the scale' or 'a semitone lower than the next highest note'.

The optional extra note that fflls the ffrst gap in our scale above is cfuarlltehde re istuhbedr iAvi dsheadr ipf (rweqruititreend ; fi) s ourc hB sfimata (lwlerri ftrteanct ffi)o. nNs oatrees rmefaeyr rbeed to as microtones.

In writing a piece of music there is usually a note that matters more than the others—the key note, or tonic.

Usually, this note is the one on which a song begins and ends. In ‘Somewhere Over the Rainbow’, the key note is the one sung on ‘Some’ and ‘I’. The way Jvuerdsyio Gna irsl asnaidd stion gbse iitn i nth teh ke effyl mof, Eth fiisa tis. E fiat (Eff). Hence, her

How is a scale chosen? The main consideration is the mood of the piece: if the composer is writing a conffdent, celebratory, up-beat piece, (s)he is likely to choose a major scale. To write a piece which is less emotionally clear a minor key is usually chosen. The primary objective diflerence between major and minor scales is that a (natural) major scale consists of a tonic (which can be any note), plus the six notes which ‘fft best’ with it, in the sense that their wavelengths bear the maximum number of the simplest possible whole-number ratios with each other. The sound waves hence combine to give the most regular/even patterns possible. In practice, one takes the pentatonic and adds two more notes. One of these two notes, the leading note, has a frequency one and seven-eighths that of the lowest note.

In a natural minor scale, two of the notes of the pentatonic, plus the leading note, have been lost (they are down-shifted by a semitone). The result is that there is no longer a big set of simple whole-number ratios, and the impact is vaguer, more complex, and less complete-sounding.

Additionally, when we hear any sung or played note, it is invariably (except in the case of some electronic

instruments) accompanied by its full set of harmonies. Major scales contain all such harmonies too, but minor scales do not. Hence, there is a genuine 'naturalness' or 'completeness' in major scales which is lacking in minor ones. 'Amazing Grace' and 'My Girl' are both in major scales, 'I Shot the Sherifl' is minor.

It used to be (and occasionally still is) claimed that individual keys have their own emotional qualities (in addition to the distinction between major and minor). Numerous experiments have demonstrated that this is not so. It is likely that this idea arose through the reputations of famous or infiuential pieces of music—imitations tended to be in the same key, which hence became associated with the emotional quality of the original piece. That is not to say that the choice of key is arbitrary: some instruments are designed so that the notes of a particular key are physically easier to play.

Sound sequences

While the raw materials of musical composition are similar the world over, there has been a great diflerence in its development: in Western music, relatively few notes are used, but they are often played together (chords) or in relation to each other (harmonies). If we had to juggle more than seven distinct notes, musical composition would be more difflcult. in particular, polyphony—the simultaneous playing or singing of diflerent melodies—would become well-nigh

impossible. In Eastern music, many more notes tend to be used, but combinations are rarer. This diflerence in approach fed through to instrument design: Western instruments like the guitar and organ are naturally polyphonic, while Eastern ones, like the sitar, can often only produce one note at a time, but can easily vary that note rapidly and subtly.

The diflerences between Western and other music may have their origins in the 10th century, when polyphony began in Europe (though there are some who claim that polyphony has far earlier roots). From polyphony sprang both the characteristically multi-tune approach of the West and the studies of harmony required to develop it. It has been suggested that the reason for the ffrst experiments with polyphony was a desire to have men, women, and children singing together.

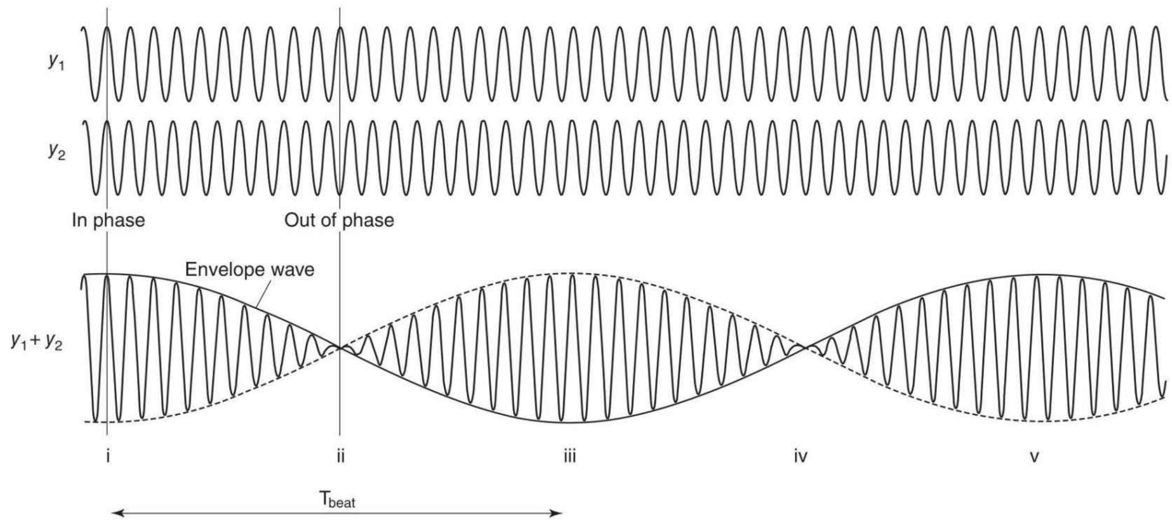
Melodies

Play some notes in sequence, and you have a tune—technically a melody. If we forget about the diflerent pitches, then what is loosely called the rhythm is all that remains—the relative lengths of the notes and rests, the speed (tempo) with which the whole thing is played, and the pattern of beats (metre, or sometimes measure). This is not quite the technical meaning of the word rhythm, which refers only to the length pattern. In ‘Somewhere Over the Rainbow’, the rhythm (technical version) is: ‘long, long, medium, short, short, medium, medium’.

The metre is indicated by the time signature. The time signature of the commonest metre (‘common time’) is written 44. The lower number gives the units used to specify the beat: 1/4 notes (crotchets). The upper speciffes the number of these beats per group. In a score, the end of each group being shown by a vertical line called a bar (so common time is ‘four crotchet beats to the bar’).

These are not the only kinds of beats that matter in music—the other kind occurs when two tones of similar wavelength play together. Every so often the peaks or troughs of the two sound waves will coincide, adding together to make regions of extra-high or -low pressure. The eflect is heard as a low-frequency ‘beat’, the frequency of which is the diflerence between the frequencies that make it up (Figure 13).

Watch any black and white thriller featuring London policemen, and you’ ll hear an example of beats when one of them blows his whistle: such whistles used two pipes of slightly diflerent lengths to produce their arresting sound. But beats aren’ t always unpleasant: some organs include a stop called the voix céleste, which controls a set of slightly out-of-tune pipes. When played along with normal pipes, the beats produced make a pleasantly wavering sound.



13. Beats.

Instrumental break

The conventional way to classify musical instruments is partly by what they were originally made of, hence strings, brass, woodwind, and miscellaneous (percussion). Another way is by whether an instrument is pitched (like a piano) or unpitched (like maracas). Any piece of music needs rhythm, and using a non-pitched source avoids dissonances with the ‘melody’ instruments. It is not essential that the rhythm source be a musical instrument: Jean-Baptiste Lully (1632-87) used to thump a big stick on the ground. This is no longer fashionable, perhaps partly because Lully died from it, having hit himself on the toe and contracted terminal gangrene as a result. Drums soon became a more popular source, though they are also used as a source of the ‘actual’ music. For this reason, while drums such as bass

drums and snare drums are unpitched, kettle drums are pitched.

Most pitched instruments make their notes either by striking, plucking, or bowing a string, wire, plate, or other solid object, or making a column of air vibrate by providing it with energy from lungs or bellows. In every case, it is resonance plus overtones (and, in the case of bells, undertones, also called subharmonics) that provide the required sound.

This makes it seem simple to construct a musical instrument, but there are three complicating factors. The ffrst is that an object such as a metal plate will not only resonate at a wavelength twice its length, but also at wavelengths twice its width and thickness—and it will produce other sounds as it twists, too.

The second point—a very useful one for string players—is that when an object such as a 120 cm string is struck (or plucked or bowed) it will move in a pattern determined not only by its length but also by the position of the striking point, since that point has no option but to be an antinode. The ends of the string have no choice but to be nodes, so if the string is struck at a point 40 cm from an end, it will make a fundamental with a node-antinode distance of 40 cm, and hence a wavelength of 160 cm. There will be another sound with a node-antinode distance of 80 cm. These components will fade rapidly as the string’ s vibration shifts towards its

natural mode of vibration, which is to have an antinode at the centre and hence make a 240 cm wavelength tone. The subjective eflect is that near-centre plucking produces a mellower sound, near-end plucking a harsher one.

Pipes and tubes behave in a somewhat similar way. In an open pipe, the simplest motion is one in which there is free movement at the open end and none (since there can be none) at the closed end: so a 12 cm pipe will have a fundamental with a node at 0 cm and the ffrst antinode at 12 cm (plus a bit due to the end correction), hence a wavelength of about 48 cm. But, just as in a string, this is not the end of the story: though the fundamental is deffned by the pipe’s length, other resonances will be produced according to its width. The mixing of these with the fundamental and its overtones produce diflerent timbres: in wider tubes, there is more energy in lower harmonics and the timbre is more rounded, while thinner pipes are brighter (or shriller, depending on your taste).The ffnal consideration for instrument makers is that a string held between two points makes a very quiet sound indeed, because its very thin cross section only aflects a tiny surrounding volume of air. In order to turn the string into a musical instrument, a much greater volume of air must be put in motion. In an electric guitar, this is accomplished by using the vibration of the metal string to induce a small electric current in pickups (magnets wound with ffne wire). In acoustic guitars and other stringed instruments, the wooden soundboard (top) is vibrated by the string and its large area moves much more air than the

string alone. Also, at low frequencies the hollow bodies of such instruments act as Helmholtz resonators.

Non-electric instruments do not add energy. Rather, they convert the energy provided by bow, plectrum, or air blast to an acoustic form. In cases where the body of the instrument shifts the frequency towards the 4 kHz region, this will sound louder, despite the fact that the sound power is actually reduced a little by frictional losses. This loudness increase stems from the construction of our hearing systems, as we shall see in Chapter 4.

Chapter 4 Hearing sound

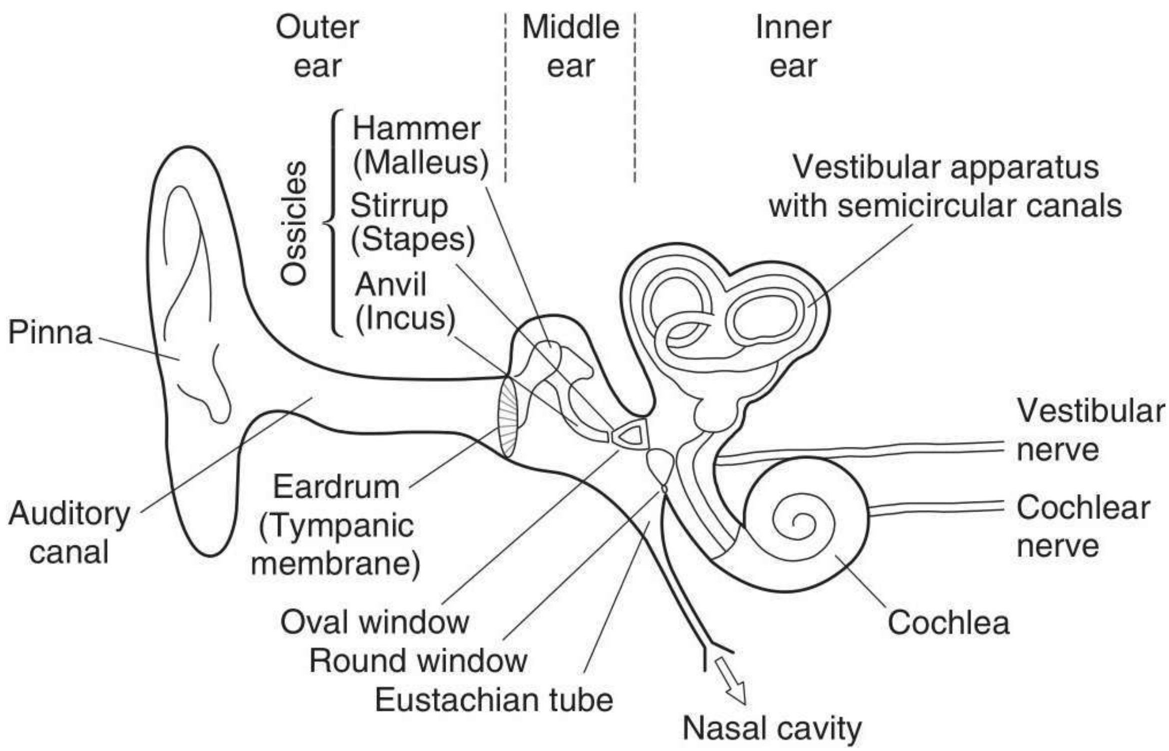
The range of hearing

Being able to hear is unremarkable: powerful sounds shake the body and can be detected even by single-celled organisms. But being able to hear as well as we do is little short of miraculous: we can quite easily detect a sound which delivers a power of 10fi15 watts to the eardrums, despite the fact that it moves them only a fraction of the width of a hydrogen atom.

Almost as impressive is the range of sound powers we can hear. The gap between the quietest audible sound level (the threshold of hearing, 0 dB) to the threshold of pain (around 130 dB) is huge: 130 dB is 1013, which is the number of pence in a hundred billion pounds.

We can also hear a fairly wide range of frequencies; about ten octaves, a couple more than a piano keyboard. We can detect, though not truly hear, frequencies well below and above this too, as will be explained in Chapter 6. And our frequency discrimination is excellent: most of us can detect differences of about a quarter of a semitone; with practice and in ideal conditions, a difference of about one-twentieth of a semitone is just distinguishable. Our judgement of directionality, by contrast, is mediocre; even in favourable conditions we can only determine the direction of a sound’s

source within about 10° horizontally or 20° vertically; many other animals can do very much better.

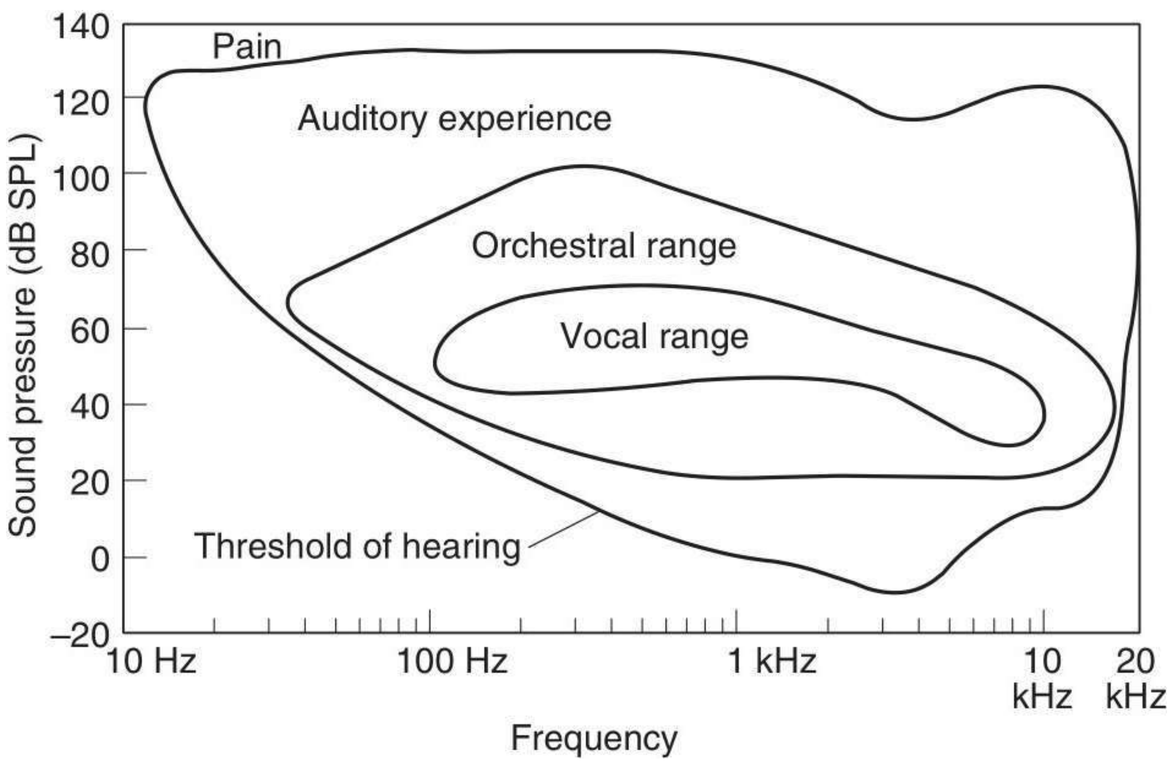


14. The ear (the middle and inner ears are greatly enlarged).

Perhaps the most impressive of all our hearing abilities is that we can understand words whose levels are less than 10 per cent of that of background noise level (if that background is a broad spread of frequencies): this far surpasses any machine.

Our ears have two functions, hearing and balance, and balance is taken care of entirely by the semicircular canals (see Figure 14). The rest of the ear has thus been free to evolve the best possible system for our needs. Not so our vocal apparatus; as a relative latecomer in our evolution it

had to flt itself in cheek by jowl, sharing structures previously earmarked for breathing and eating, licking and sucking, kissing and flghting. Yet, in a trained actor or accomplished singer, the speech system functions just as perfectly as a Stradivarius violin.



15. Frequency dependence of hearing.

What do ears hear?

As Figure 15 shows, our hearing systems have not evolved to measure the physical powers of sounds: a piccolo, for example, has a maximum power of about 0.08 watt, while a trombone can manage about 6 watts, yet it sounds the quieter instrument. This means that trombone players must work a lot

harder than piccolo players (who in fact never play at full blast in any ordinary piece of music). But all instrumental powers are low. A loud orchestra playing at full stretch might manage 60 watts; if it could keep it up for two minutes, that would be enough to boil a tablespoon of water.

Moreover, this power spreads from the source in many directions, so, if you were sitting 10 metres from such an orchestra, less than 0.01 per cent would reach your eardrums. And what they actually detect is nothing more than a series of rapid prods of air, which carry no information other than how hard they hit and how rapidly they arrive. That we can experience a whole world of sound, charged with emotion and packed with meaning, is due to the precise coordination of highly evolved anatomical, electrochemical, and neurological processing systems.

The outer ear: capturing the sounds

The pinna, the visible part of the ear, acts mainly as a funnel to collect sounds; its front-to-back asymmetry also provides some information about their direction. The moveable pinnae that some mammals possess greatly assist in direction flnding; though some of us also have this ability too, it is of no beneflt beyond its entertainment value.

Since the auditory canal is a cylinder around 30 mm long, it has a resonance at a wavelength twice this, corresponding to a frequency around 3 kHz. The gain in energy at this

frequency results in a loss at others, so the canal acts as a partial band-pass fllter.

At the end of the canal is the eardrum (tympanic membrane), a roughly circular disc of stretched skin about 1 cm across. It is inclined at an angle to the canal, to maximize its area and so capture as much force as possible. It is also slightly conical, which allows it to transfer more energy than it could if it were fflat. The drum has no resonances—for frequencies above ~3 kHz, its surface moves in a chaotic way, and for lower frequencies it all moves as one piece, since the waves are larger than it is. As a result it transfers the widest possible range of frequencies with minimal flltering. This near-fflat frequency response is achieved in part through its asymmetric shape and in part through an internal scaffold of collagen flbres.

It is important that the drum is taut but not stiff, and this is made possible by the equalization of internal and external pressure by the Eustachian tube (always drawn as open despite the fact that it never is, except for an instant when there is a significant pressure change—making a distinctive crackling sound when it operates).

The middle ear: sharpening the blow

The eardrum is connected to a series of three tiny (tiniest, in fact) bones called ossicles, which occupy the air-fllled middle ear. Their main job is to covert the wide shallow motion of the eardrum to a higher-pressure ‘tap’ on

a second drum-like membrane called the round window, which is the gateway to the inner ear. By working as levers, the ossicles increase the force by about 1.5 times. However, the main way in which the force is enhanced is simply through the ratio of the areas of the eardrum and the round window, which increases the pressure about twenty times by concentrating the force over a smaller area. The ossicles also provide some protection for the inner ear, through the acoustic refflex (see Chapter 8).

The inner ear: sound to electricity

The inner ear is full of liquid, and, just as the eardrum converts airborne sound to bone-borne, so the round window converts the latter to a ffluid-borne version. This passes along the cochlea, a coiled tube about 2 cm long. At its extremity is a hole (the helicotrema), and the wave passes through this and then travels back along a second tube which is joined along its length to the first. When the wave has completed its doubled-back journey, it must be eliminated, otherwise it would refflect back up the cochlea and interfere with newly arrived waves. So, the second tube terminates in yet another membrane, the oval window. This bulges outwards when the waves reach it, dissipating them as heat.

Running between the tubes like the fllling in a baguette is the basilar membrane, which converts the waves to nerve impulses.The side-by-side tubes would be about 5 cm long, were they not coiled up like a snail (cochlea is Latin for

snail). Because the length of the cochlea is related to the wavelength of sound waves, it is similar in all mammals: only about 50 per cent longer in an elephant than a person, so the coiling is probably simply a space-saving measure, rather than having any acoustic function. Mice and other tiny mammals cannot accommodate a full-size cochlea. Theirs are about 1 cm long—and consequently they can only hear about three or four octaves compared to the eight to ten octave range which we and most other large animals can detect. On the membrane is mounted the organ of Corti, on which grow around nine rows of short hairs called stereocilia (around 400 per row). These rows run the length of the membrane, and have nerve flbres attached to them. These flbres bundle together to form the auditory (cochlear) nerve, which transmits impulses to the brain.

The basilar membrane moves in response to the sound waves that impinge on it. It is stiffer and wider at its root than at its tip, which means that lower frequency sounds cause oscillations closer to the latter. These oscillations cause the stereocilia to move, and the hair cells to which the stereocilia are attached then send electrochemical impulses to the brain. Since the brain knows which cells are where, it can determine the frequency of sounds by this means (this is known as place theory). However, when presented with sounds below about 1 kHz, the whole membrane oscillates. In this case, a different mechanism becomes important,

in which the hair cells flre in time with the pulses that make up the sound wave—flring one hundred times a second in response to a 100 Hz tone for instance.

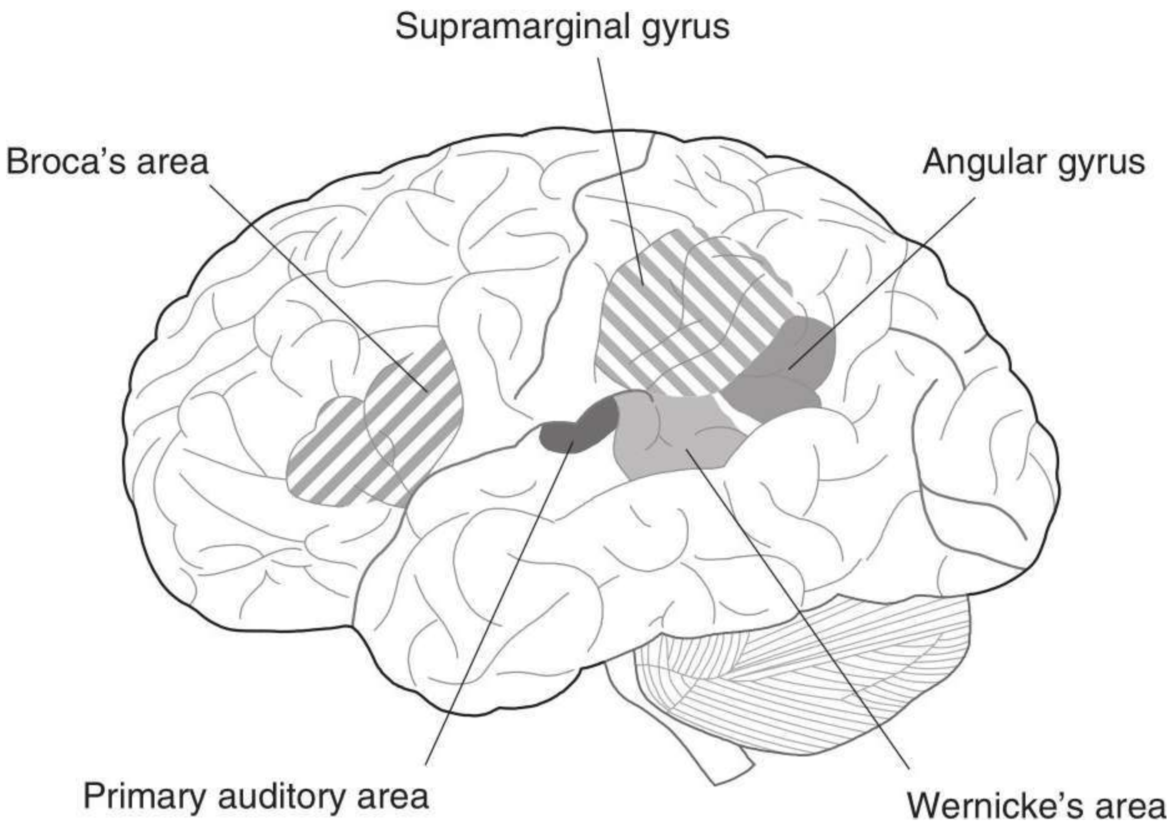
However, the cells are incapable of flring more rapidly than about 500 times a second. To respond to higher frequencies they therefore face a similar problem to that of a squad of soldiers armed with fflintlocks which take, say, a minute to reload: how can the squad produce sustained gunflre at intervals of 10 seconds? The answer is to divide the squad into six groups. The flrst group flres and starts to reload. Ten seconds later, the second group flres, and so on. Ten seconds after the sixth group has flred, the flrst group will have completed its reloading and will flre again. Hair cells work in groups in just this way: the flrst group might signal the brain when a cycle of the sound wave is at its peak, the second when that cycle has fallen halfway to its minimum, the third at minimum and the fourth when it is halfway back up to maximum. In this way, such a quartet of cell-groups could respond to a tone four times higher in frequency than their individual maximum flring rates.

The hair cells of the basilar membrane also work together in a very different way: those in the eight or so outer rows respond to incoming sound waves by changing their lengths in time with them. This motion amplifles the vibrations of the stereocilia on the inner row of hair cells (the only ones that send signals to the brain) and hence provides signiflcant ampliflcation, allowing us to hear sounds which

are 40 dB (one ten-thousandth) less powerful than we otherwise could. This activity generates its own faint sounds, called otoacoustic emissions. These emissions are of great value in determining the functionality of the hearing system in infants who are too young to report whether or what they can hear. Also, they often fade when there is any damage to the inner ear, so they are a useful check for audiologists (hearing specialists) too. They are (fortunately) far too quiet to be heard, so sensitive in-ear microphones are used to measure them.

Nerves and brain: objective to subjective

The nerve signals that emerge from the basilar membrane are not mimics of sound waves, but coded messages which contain three pieces of information: (a) how many nerve flbres are signalling at once, (b) how far along the basilar membrane those flbres are, and (c) how long the interval is between bursts of flbre signals. The brain extracts loudness information from a combination of (a) and (c), and pitch information from (b) and (c).



16. Locations of hearing, language, and speech activities in the brain. The primary auditory area recognizes sounds. Broca' s area both analyses and produces semantic elements. Wernicke' s area deals with the sequences of speech sounds (heard, made, and remembered). The supramarginal gyrus may deal with articulation, and the angular gyrus may assist with semantic processing.

What happens in the brain' s hearing and language centres (see Figure 16) is not entirely clear, but the flrst stage in processing is to extract salient features from the stream of input data from the auditory nerve. These features are then used to continuously update, amend, and reflne a mental model of the thing being listened to—perhaps a tune, a spoken phrase, or a worrying engine noise. The model' s accuracy is tested by predicting what the sound will do next. What the

brain is trying to do is establish the degree to which each component of a sound contributes to the meaning, a process called hierarchical encoding. Our incredible prowess in performing this function is demonstrated by our ability to, for example, recognize that someone is trying to hum 'Somewhere Over the Rainbow' even if they get most of it wrong. The speed with which the mammalian brain interprets the sounds that reach the ears is almost incredible—say 'squirrel' to a dog who knows the word, and the response is almost instantaneous.

Homing in

Following a piece of music is a rather unusual and newfangled (in evolutionary terms) thing for a brain to do: usually the focus of interest is an object in the external world, and consequently an important function of the hearing system is to locate that object. For high frequencies, the brain relies partly on the fact that a sound coming from the left will arrive flrst at the left ear, and partly on the blocking effect of the head, which means that the level at the right ear will be lower. For a sound wave longer than the distance between the ears, the brain compares the changing levels of sound at the ears: if a long wave arrives from the left, each of its antinodes will reach the left ear flrst, so the pressure at that ear will initially be highest. As the wave progresses, the left ear pressure falls while the right rises until the antinode passes it, when it will begin to fall again. However, with a wave more than about 4 metres

long, there is very little change in level over the inter-ear distance, so its direction cannot be judged.

To determine whether the sound source is above or below the ears, the brain relies on the effects of the shapes of the head and shoulders on the level of the sound. Direction flnding is not the only advantage of having two ears: the auditory nerves sometimes flre even when no sound is present, but the brain will reject such signals if they come only from one side.

The brain's processing system has evolved to make reasonable assumptions about the sounds it receives, leading to such phenomena as the precedence effect (also known as the law of the flrst wave front or the Haas effect). The brain's assumption is that a sound that arrives in the flrst fraction of a millisecond indicates the direction of the sound source. So, subsequent sounds are regarded as coming from the same direction as the flrst. This allows us to locate a sound source in dark spaces without being confused by echoes arriving from many directions. Such assumptions can mislead, especially in situations that would not occur naturally: for instance, if one listens to a sound from a loudspeaker about a metre away and at 45° to the left, and this sound is gradually replaced by an identical one from a loudspeaker at 45° to the right, the sound will still seem to come from the left.

A very useful feature for people trying to communicate in crowds is the cocktail party effect, in which particular phrases (such as one’ s name) stand out from the hubbub. This works for non-vocal sounds too: conductors are often highly attuned to particular instruments or musical phrases. This effect works because the brain is constantly model-building whether we are actively listening or not, and because it preferentially seeks matches with sounds that it has classifled as having a signiflcant meaning, like its owner’ s name.

The role of hearing is a very rich and complex one—as linguist Roland Barthes points out, sounds act on our minds in three ways: as ‘indices’ (the alarming sound of an explosion), ‘signs’ (the literal meaning of a word), and ‘signiflers’ (unconscious associations triggered by a word like ‘end’). And hearing is also very much a social activity too—according to Labelle:

the rich undulations of auditory material do much to unflx delineations between the private and the public. Sound operates by forming links, groupings and conjunctions that accentuate individual identity as a relational project . . . [and] weave an individual into a larger social fabric . . . contributing to the meaning of shared spaces.

Labelle points out, too, that whether we like it or not, sounds bring us into intimate contact with others: crying babies, noisy neighbours, or cheering fellow football

supporters. As he puts it: ‘Sound creates a relational geography that is most often emotional, contentious, ffluid.’

Hearing bones

The eardrums are wonders of evolutionary engineering, but we can actually hear fairly well without them, since sound waves also reach the inner ear by travelling through the bones of the head, speciflcally the mastoid bone behind the ear. Submerging one’ s ears in the bath largely switches off the airborne route so the sounds that remain arrive mainly by bone conduction. This system is rather insensitive however: using air conduction, we can hear sounds about 40 dB less powerful than the weakest detectable through our mastoids. On the other hand, bone conduction allows us to detect sounds with frequencies up to 30 kHz, which is well beyond the upper frequency limit of the airborne route—but, presumably because such sounds are of little value to us, they are all encoded in the same way, so they give rise to the same pitch sensation as 20 kHz sounds.

Bottlenose dolphins (Tursiops truncatus) take bone-based hearing much further: their jaws bear teeth that are spaced at regular intervals and set at the same angle, are all of a very similar shape, and have heights that depend upon their location. This adds up to a focussing array, in which sound waves of particular wavelengths are signiflcantly amplifled—but only if the source is directly ahead. Thus, the dolphins

can hear very quiet sounds, and can determine their directions simply by moving their heads until the loudness is maximized.

Deafness

The hearing system is a delicate one, and severe damage to the eardrums or ossicles is not uncommon. When it occurs we must rely instead on bone conduction and artiflcial aids: Edison used his teeth to transfer the sounds from his phonograph to his mastoid, as the marks on the device still testify.

This condition is called conductive hearing loss. If damage to the inner ear or auditory nerve occurs, the result is sensorineural or ‘nerve’ hearing loss. It mostly affects higher frequencies and quieter sounds; in mild forms, it gives rise to a condition called recruitment, in which there is a sudden jump in the ‘hearability’ of sounds. A person suffering from recruitment and exposed to a sound of gradually increasing level can at flrst detect nothing and then suddenly hears the sound, which seems particularly loud. Hence the ‘there’ s no need to shout’ protest in response to those who raise their voices just a little to make themselves heard on a second attempt.

Sensorineural hearing loss is the commonest type, and its commonest cause is physical damage infflicted on the hair cells. With very high levels of sound, the eardrums can be ruptured (and they can also be damaged by blows to the head

or by infection). Remarkably however, burst eardrums can not only heal, they usually then perform almost as well as before. About 360 million people worldwide (over 5 per cent of the global population) have 'disabling' hearing loss—that is, hearing loss greater than 40 dB in the better-hearing ear in adults and a hearing loss greater than 30 dB in the better-hearing ear in children (who make up about 10 per cent of the total). About one in three people over the age of sixty-flve suffer from such hearing loss.

In terms of alleviating deafness, what can be done varies greatly. For conductive hearing loss, if some hearing function remains, hearing aids tailor-made to flt the wearer's particular pattern of loss are highly effective. They may also be equipped with noise-cancelling functions and often use directional microphones, so that the user can focus on sources that they are looking at. They can also include a vibrating element to stimulate the mastoid. Hearing aids are less successful in treating sensorineural hearing loss, because signals are distorted by recruitment.

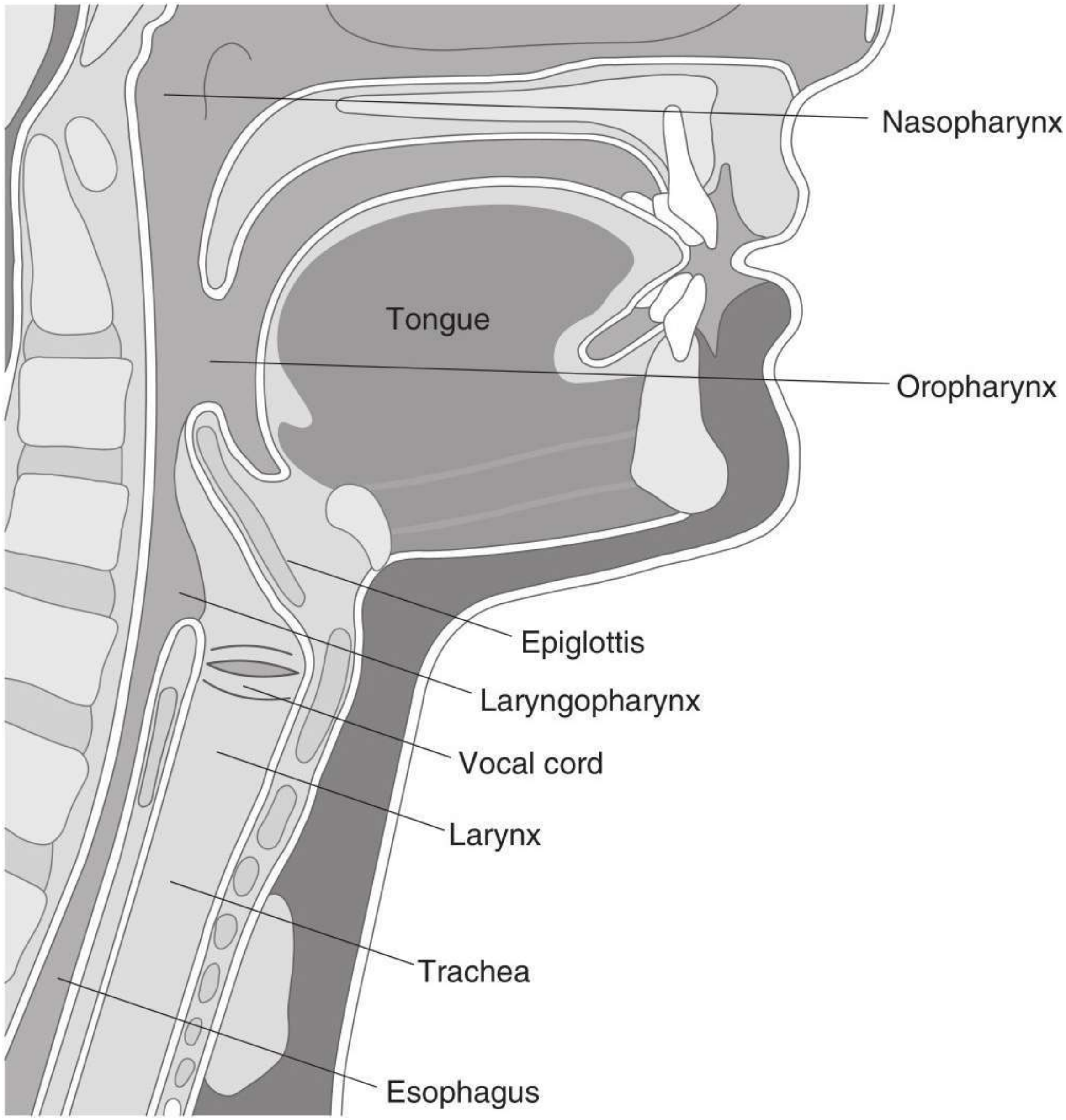
With complete deafness the challenge is much greater, but in the last few years the introduction of fairly reliable cochlear implants has given new hope. And the future looks promising: in 2012, by encouraging stem cells to grow into hair cells, deafened gerbils had their hearing restored, typically by 45 per cent, but by 90 per cent in a few cases. This approach may one day be applicable to those people (around 10 per cent) whose hearing loss is caused by damage

to nerves called spiral ganglion neurons. There are also some animals, including owls, in which lost stereocilia simply grow back, and it may be that this capability could be genetically induced in humans (without adopting the other peculiar feature of owl hearing, which is that it only works at its best in spring, presumably because owls need to catch extra prey for their chicks then).

There are a great many hearing problems which are not simply characterized by loss of function. The commonest is tinnitus: ringing in the ears. Its causes are largely mysterious, and the level, type, and duration of the sounds vary greatly from person to person. It is often associated with past infection, drugs (especially certain antibiotics), or trauma, and it frequently accompanies hearing loss.

Structures for speech

Around a million years ago, the hearing systems of our ancestors underwent subtle changes to flne-tune them—literally—to detect speech. We know little about the evolution of our vocal systems; although the ability to make sounds in a controlled way is common to most animal species, speech is immeasurably more complex. So, unlike the evolution of, say, the leg, we cannot look back at a long chain of ancestral forms and watch the system adapt to the changing requirements of those users and the shifting demands of their environments.



17. Vocal apparatus.

In its most basic form, the making of sounds is simple: Figure 17 shows the structures involved. Air exits from the lungs through a tube—the windpipe or trachea—which is equipped with fflap-like vocal folds (vocal cords) that restrict the air fflow and vibrate when they are tensed by

muscles attached to them. Increasing this tension increases the vibration frequency, but the length of the folds sets its lower limit, resulting in fundamental frequencies round 125 Hz for men, 200 Hz for women, and 300 Hz or above for children. In boys, there is a sudden increase in length during puberty, which causes the ‘breaking’ of the voice.

After emerging from the end of the trachea, this low-frequency sound enters the back area of the vocal tract, roofed by the soft palate. In front of this is the hard palate. Together these structures form a cavity in which the sound forms resonances called formants. The characteristic wavelengths of vowels are set here, and varied by raising the tongue to change the volume of the tract or to divide it into two linked cavities.

Consonants involve more parts of the vocal apparatus than vowels, are usually shorter in duration, and, in many cases, change while they are being made. There are four main types, deflned by their manner of articulation.

Plosives are made by the sudden stopping of airfflow (hence their alternative name: stops). Fricatives and liquids require partial stopping, with or without turbulence respectively. Nasals defflect the airstream to the nasal cavity. Glides (semivowels) involve a rapid transition from one vowel sound to another. The unusual helpfulness of this naming system is shared by the subdivision of the consonants according to their place of articulation, as Table 4 shows.

Also shown is whether or not the consonant is voiced—that is, whether the vocal cords are involved in making it. (Ventriloquists attempting to produce labial or labiodental sounds are stymied by the need to keep their lips separated and motionless. Skilled proponents of the art circumvent this by speaking very fast indeed at such points.)

Since the wavelengths of the resonances in the vocal tract depend only on its structure, changing the velocity of sound alters the frequencies of those wavelengths. Hence the 'Donald Duck' voice produced by breathing helium (14 per cent of the density of air), and the more rarely heard gravelly voice produced after breathing xenon, which is 4.6 times denser than air.

Manner of articulation		**Place of articulation**						
		Labial (lips)	**Labio dental (lips and teeth)**	**Dental (teeth)**	**Alveolar (gums)**	**Palatal (hard palate)**	**Velar (soft palate)**	**Glottal (glottis)**
Stop	Voiced	b			d		g	
	Unvoiced	p			t		k	
Fricative	Voiced		v	th(en)	z	zh		
	Unvoiced		f	th(in)	s	sh		h
Nasal	Voiced	m			n		ng	
Glide		w			y			
Liquid					l & r			

Table 4. English consonants

Hearing sound

Life would be a dull thing, however, if all we did with our vocal apparatus was speak. Singing is, physiologically, no different from speaking except that every aspect of the

sound made is more flnely controlled, and pitch is often keyed to an externally deflned value. Whistling does not involve the vocal cords: it requires the production of turbulence around the lips, which transfers energy to the vocal cavity, which acts as a Helmholtz resonator. Shouting simply requires greater air force from the lungs. In whispering, the vocal apparatus works as it does when producing normal speech, except that the vocal folds are neither vibrated nor fully relaxed, so that when air passes through them it produces turbulence (this is called adduction). Since much more air can pass between the folds without exciting sound waves, whispering is necessarily relatively quiet.

Our hearing systems are far more sophisticated than our most advanced machines, and have evolved to suit us admirably. But what nature has given us is limited in range. In all but the tiniest groups of people, communication—which we prize so highly—must spread far beyond the reach of voice or of hearing, and it was to answer that need that flrst electricity and then electronics were pressed into service, to augment and to replace our fflesh and our nerves. How they do this is the subject of Chapter 5.

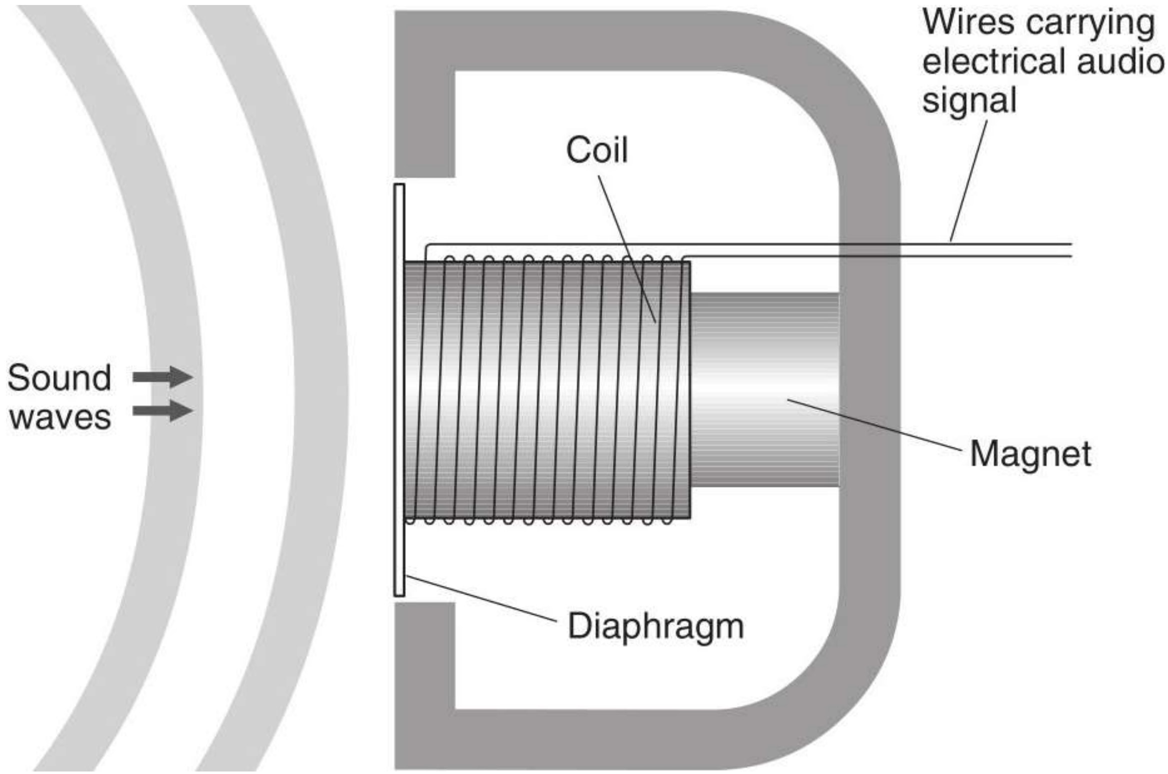
Chapter 5 Electronic sound

Sound to electricity: microphones

As invented by Charles Wheatstone in the 1820s, the microphone was a purely acoustical device made of two metal plates clamped to the ears by a springy rod and a length of dressmaker’s ribbon. To use it, one pressed one’s head against the sound source (Wheatstone helpfully suggested a boiling kettle) which, with a bit of luck, could be heard more clearly. To no one’s surprise but Wheatstone’s, it didn’t catch on.

The ancestor of today’s devices—which convert sounds to electricity rather than simply directing them to the ear as Wheatstone’s version did—was the carbon microphone, invented by David Hughes in the 1870s. In this, a thin metal plate (the diaphragm) compresses a vessel filled with carbon granules through which a current is passed, and the compressive force alters the electrical resistance. Although their performance was very poor, carbon microphones were used in telephones for decades.

Although there have been many microphone designs over the years, and several specialist types are available today, only three major ones are now in common use: the dynamic or moving coil microphone, the condenser microphone, and the piezoelectric microphone.

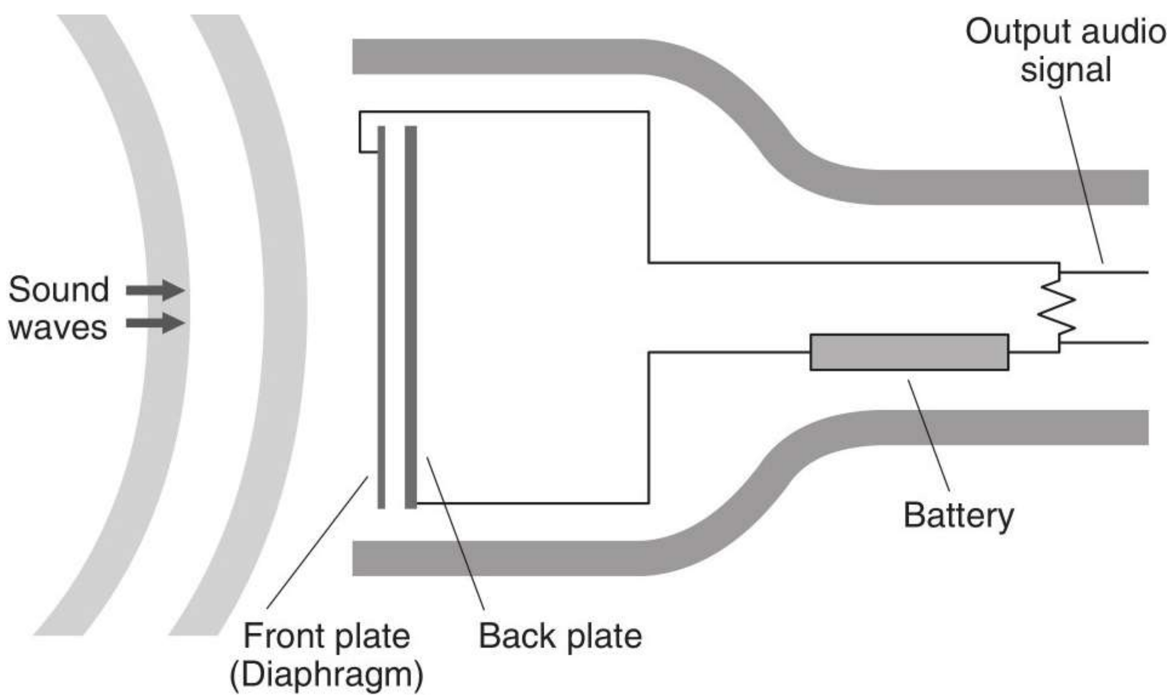


18. Dynamic microphone.

In a dynamic microphone (Figure 18) the diaphragm is attached to a coil of wire that surrounds a stationary magnet. A small voltage is induced in the coil when it moves, which in turn gives rise to a weak current. The low voltage means that the quality of the response is too poor for measurement work, so dynamic microphones are mainly to be found in concerts and recording studios.

In a condenser microphone (Figure 19), the diaphragm forms one plate of a capacitor (formerly called a condenser, hence the name). A capacitor is a pair of parallel metal plates with a thin layer of either air or some other material which does not conduct electricity (known as a dielectric) between

them. Attaching one plate to the negative pole of a battery charges it with electrons. All metals contain free electrons, and those in the other plate are repelled by the electrical field arising from the large number now residing in the negative one. The repelled electrons ffow from that plate, to leave it positively charged. The whole capacitor, therefore, now has a voltage (called a polarizing voltage) across it, and the microphone is ready for use: the diaphragm plate moves in and out with the compressions and rarefactions of the sound waves that impinge on it. A large electrical resistance stops the charge from rapidly escaping, so instead the sound waves are transformed to patterns of voltage ripples. Condenser microphones have an excellent frequency response and are used as measurement microphones in laboratories and in SLMs. They also respond more rapidly to sudden sounds than do dynamic microphones.



19. Condenser microphone.

Crystal microphones and ceramic microphones exploit the piezoelectric eflect, in which quartz or some other crystalline material produces a voltage when slightly compressed. Most landline telephones use these, as do call-centre headsets.

Other types of microphone, of historical or specialist interest, include the following:

Electret microphones are made of permanently charged material (the electrical equivalent of a magnet). They function in a similar way to condenser microphones.

MEMS (microelectromechanical systems) microphones are condenser microphones etched directly on to silicon chips.

Just a few square millimetres in area, their cheapness and robustness means they are used in mobile phones, among many other applications.

Optical microphones use a shiny silicon membrane as a diaphragm, which reffects the light from a light-emitting diode (LED). A photo detector measures changes in the light when the membrane is vibrated by sound waves, and an electronic circuit converts these changes into an electrical signal. Such microphones are compact and robust and are unaflected by local electromagnetic fields, so they are used, for example, to allow communication between patient and stafl during MRI scans.

Pressure zone microphones (PZMs) are designed for use near a hard reffective surface (such as when placement directly on a stage ffoor is required). With a conventional microphone, reffection from the surface would interfere with the direct sounds to the microphone, but a PZM overcomes this by having its diaphragm so close to the surface that most wavelengths overlap.

Ribbon microphones use a strip ('ribbon') of metal rather than a diaphragm, and unlike the eardrum and most other kinds of microphone, they respond to the pressure diflerence between the ribbon' s sides, which mean that velocities with which air molecules in the sound wave move are detected, rather than sound pressure patterns. Such microphones are often used by commentators in noisy surroundings, because a sound which comes from all directions will impact both sides of the ribbon simultaneously, so that it will not respond. For this application, the microphones have a projection that is held to the upper lip, which helps guide the commentator' s voice so that it impinges on one side of the ribbon only.An alternative for capturing speech in noisy environments is the lavalier microphone, which is a small clip-on electret or dynamic microphone. Lavaliers have the advantage that they can readily be concealed under clothing.

A sound intensity probe consists of a pair of face-to-face microphones which measure the pressures of diflerent parts of the same sound wave. From these the molecular velocity and hence the sound intensity is calculated.

An important criterion in choosing a microphone is directionality: an ideal omnidirectional microphone is equally sensitive to sound from any direction, and is required to capture a full soundscape, for example. A unidirectional microphone, which picks up sound from one direction only, is ideal for picking up speech or song in noisy environments.

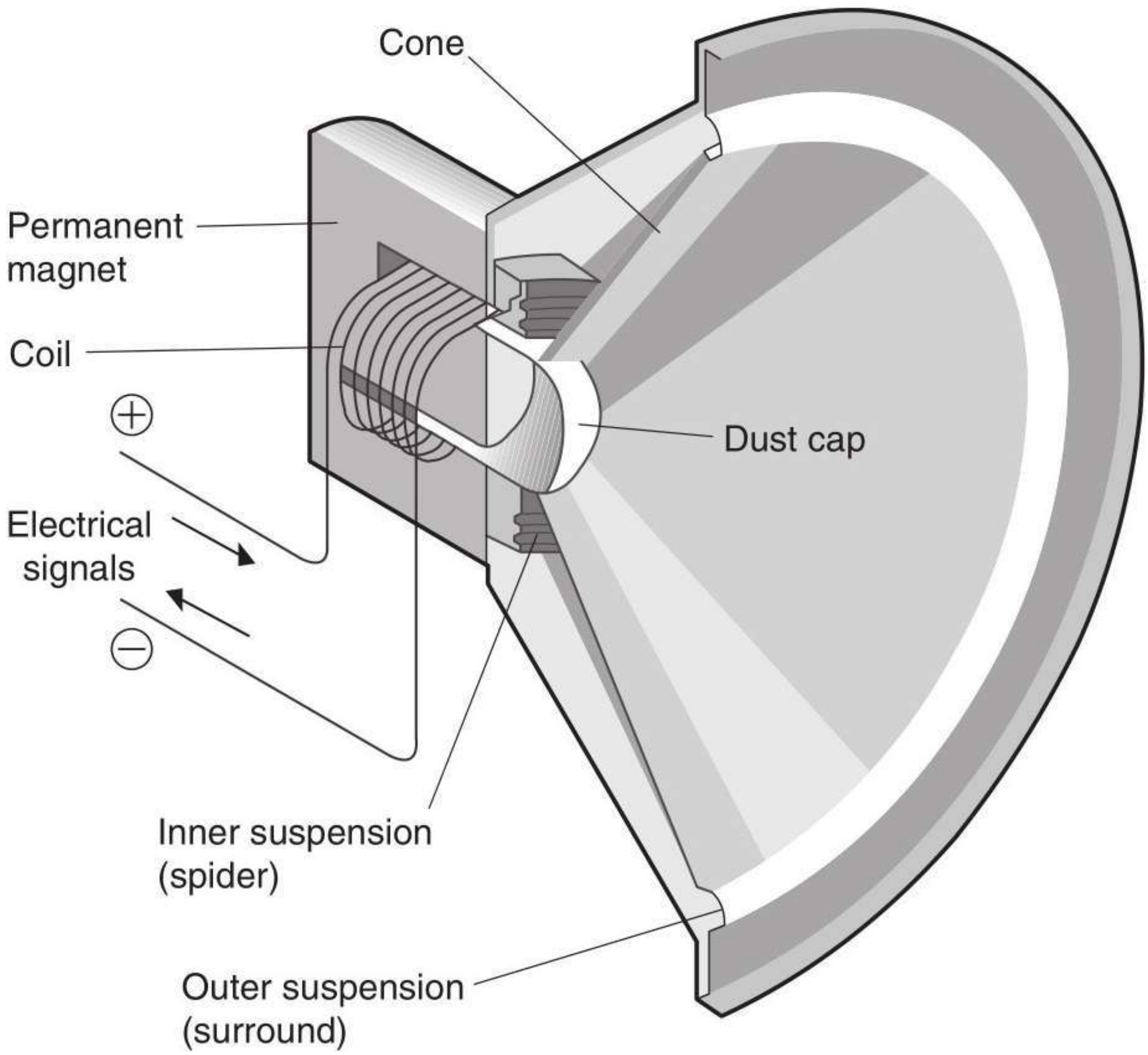
Except for the ribbon microphone and sound intensity probe, every type is naturally fairly omnidirectional in response, providing that the wavelengths of the sound waves that impinge on it are larger than the diaphragm. In order to make a microphone sensitive mainly to sound from immediately in front or behind (bidirectional), all that is needed is to allow the back as well as the front to be open to the air. The diaphragm will not move very much in response to sound waves arriving from all around it, because the rise and fall of pressure at the front would be very similar to that at the back. But any waves incident on the front or back face of the microphone will be picked up readily.

A simple way to achieve directionality is to mount a microphone at the focus of a reffector whose cross section is a parabola. This shape reffects incident sound waves on to the diaphragm, as long as they are smaller than the reffector. Alternatively, a microphone can be mounted at the end of a tube with slits along its sides, to make a shotgun (boom) microphone. Sound waves passing straight down the tube travel unimpeded to the microphone, but those from other

directions enter through the slits. Each such sound will enter multiple slits, so many versions of it will be formed, each with a diflerent phase. The versions will therefore largely cancel each other out through destructive interference. Shotgun microphones are widely used for outdoor recording, often accompanying cameras. They are, however, highly frequency dependent.

Electricity to sound: loudspeakers

Loudspeakers are microphones in reverse: if a dynamic, condenser, crystal, or ceramic microphone is supplied with a varying current, it will vibrate to produce sound waves (such microphones are therefore referred to as reciprocal transducers). Most actual loudspeakers are dynamic microphones in reverse, and are called moving coil loudspeakers (Figure 20). As the name implies, an electrical signal is fed to a coil (the voice coil) attached to a diaphragm which is usually in the form of a cone. The coil surrounds a magnet and the electromagnetic field set up in the coil by the signal causes it, and the diaphragm, to move.



20. Loudspeaker.

Loudspeakers are intrinsically inefflcient devices: most of the electrical energy that passes into them appears as heat, with only about 1 per cent being audible. Hence, amplification is vital. Transistors make amplification a simple matter today, the main challenge being to ensure that each frequency is amplified to an appropriate extent, which, given the non-linear attributes of the hearing system, means

that diflerent degrees of amplification must be applied to diflerent frequencies if the pitch envelope of the output sound is to remain unchanged. It is essential to deploy them with care. Quite apart from the deleterious eflects on hearing, loudspeakers can be easily damaged, especially by artificially produced sounds, which may have an extremely fast rise time. Also, if the microphone picks up the sound of the loudspeaker to which it is sending its output, a positive feedback loop is set up, resulting in an all too familiar squeal.

Loudspeakers are the least hi-fi link in the music production chain: though simple in principle, their design presents many practical challenges. The voice coil must return to exactly its start position when the signal falls to zero, with no oscillation beyond it, and yet must be free to move. The cone must hold its shape while it vibrates, be very light, yet rigid enough not to sag under gravity; large enough to move large volumes of air at low frequencies (where only powerful waves are audible), yet small enough so that they can move back and forth over 10,000 times a second at high frequencies. Also, the casing must not resonate at any frequency.

In practice, it is much easier to use speakers in groups, usually all in the same unit—a small tweeter for frequencies above 2 kHz, a larger mid-range speaker (50 Hz to 5 kHz), and a woofer (30 Hz to 800 Hz). For those who love lots of bass, there may also be a subwoofer (20 Hz to 200 Hz).

Subwoofers are usually active loudspeakers, which means they contain their own amplifiers (and hence need their own power supply). Most other speakers are passive, driven by a signal which has already been boosted by an amplifier within the hi-fi (or other) system.

A loudspeaker without a case is almost silent, for the simple reason that the high-pressure pulses generated at its front simply slip round to the back to fill the low-pressure area that has just formed there. Hence, loudspeakers may be mounted in an airtight box. If the box is small, however, the air in it resists being compressed as the diaphragm moves in. An alternative solution is to place the diaphragm in the centre of an annulus called a bafie. The bafie must be large enough such that, by the time the pressure pulse has travelled round it to the back of the diaphragm, the low-pressure area there has gone (or, to put it another way, the distance the sound waves must travel is longer than a quarter of a wavelength for the longest waves—the lowest frequencies—of interest).

Helmholtz resonance can be used to extend the low-frequency performance of a loudspeaker. A hole (port) is made in the front of the box in which the speaker is mounted, and the cavity then resonates at low frequencies. If the resonant frequency of the box lies below that of the speaker, a pressure pulse produced when the speaker diaphragm moves back (the backwave) will make its way within the box to emerge at the port, where it will be in phase with a new pulse just

produced by the front of the diaphragm. These two in-phase pulses will reinforce each other. This applies to pulses which make up waves of any frequency that falls between the resonance frequency of the diaphragm and that of the box.

One disadvantage is that the pufls of air from the port can sometimes be heard, and another is that sounds are less crisp because each signal is followed by a short fading ‘tail’ of resonances. Also, pulses forming sound waves with frequencies below that of the port are cancelled out by the subsequent wave from the diaphragm. Frequencies above both resonances are neither enhanced nor reduced.

Our brains are so good at filling in gaps in sounds that we can exploit this to make do with quite basic loudspeakers. In nature, a set of tones of 200 Hz, 300 Hz, 400 Hz, and 500 Hz will almost always be harmonics (overtones) of a fundamental at 100 Hz. Because the brain’s hearing centre (see Figure 16) ‘knows’ this, it will confidently decide that the 100 Hz is actually there. But, if the set of tones is coming from a small loudspeaker, it’s likely that there will not in fact be a 100 Hz tone present. This eflect is known as the missing fundamental, and is why the lack of low frequencies in the Epidaurus theatre (Chapter 1) did not sound strange: the hearing centres of the audience interpolated them. It also explains the relative clarity of the earliest telephones, which did not transmit low frequencies well.

The introduction of eflective microphones, amplifiers, and loudspeakers revolutionized the record industry—and this revolution could be rapid since the records themselves did not need to be changed. Social changes soon followed: quite suddenly, almost anyone could access music, and select what they listened to as well.

According to sound studies expert Jonathan Sterne, such new sound media ‘called into question the very basis of experience and existence’. And the eflects on performers were profound: as music historian Robert Philip points out, many were aghast at the number of errors they heard in their recorded performances. Musicologist Mark Katz suggests that such performers became trapped in a ‘feedback loop’ in which they attempted to produce more and more ‘perfect’ performances, only to be disappointed all over again when they listened to recordings of the result. As a result, performance became less individualized, more standardized, and lost spontaneity. The experience of listening to recordings caused another kind of feedback too: violin vibrato, for example, was originally a phonographic eflect, but was soon being imitated by performers.

The next revolution in this field was the invention of stereophonic phonograph recording in 1933, achieved by recording the two channels on the two walls of the groove, at 90° to each other and 45° to the vertical. The introduction of stereo records meant that in principle the whole three-dimensional sound field of the original performance could be

recreated, giving rise to questions of ideal speaker placement which have fascinated music bufls to this day. It also gave rise to the concept of 'fidelity', since what one was now doing was recreating an original performance. For a while at least—with advancing post-processing and mixing techniques, by the 1970s many pop pieces were never performed as such; much of their content was added after the band had gone home. For classical music however, faithful recording has remained key. Nevertheless, despite decades of interest by many millions of amateurs and professional listeners, recorders, performers, and players, 'fidelity' is still unquantifiable.

Storing sounds

Once the technology of amplification and loudspeaker design had been perfected, the main concern was the fragility of records. A whole culture developed concerning how to handle, house, clean, and—once correctly trained—play them. The auto-changer came as a great relief to some, though to others was regarded as either a spoilsport or an incipient record-damager. Partly because of this mystique and partly because of the high quality of some of their covers, vinyl records, and especially LPs (later called albums), were venerated as objects in a way that no other recording media have ever been, and a few are still sold to this day.

In 1964 analogue cassette tapes were introduced as a robust and compact alternative to discs, and to begin with

were very popular: although recorded ones were readily available, many preferred to buy a vinyl record and (illegally!) transfer it to cassette to listen to while keeping the record pristine. Radio broadcasts could also be recorded in this way and the music centre, comprising radio, cassette recorder, and record player, became popular since it allowed cassettes to be recorded with a minimum of fuss.

There are two major disadvantages with tape, however: track selection requires time-consuming winding, and high-frequency hiss is unavoidable. The latter was reduced somewhat by the many variant Dolby systems on ofler, all of which work by recording a version of the track with the high frequencies boosted, and then suppressing them on playback. This is an example of a technique called companding (a portmanteau of compressing and expanding).

Just as stereo spawned new interests and behaviours, so did tape. Mixer tapes were one outcome, and another was the Sony Walkman, which at last allowed music lovers to listen to their music wherever they happened to be, while only annoying people sitting next to them.

But what was needed was to do away entirely with analogue recordings, that is, those in which a sound is stored as a continuous varying pattern (whether physical, as in a vinyl record,

or magnetic, as in a cassette tape). In a digital system, the recorded signal is encoded as a string of numbers, and

once in that form it can be stored, transmitted, or copied with neither degradation of the signal nor increase in background noise.

At first glance, it may seem that in order to capture the intricacies of a complicated sound wave (like those in Figure 9), the amplitudes of a great many points of that wave must be measured and coded. In fact, the sound need only be sampled twice as frequently as the highest-frequency component that one wishes to preserve. So, to encode all frequencies in a signal up to 8 kHz, one must sample at 16 kHz (this is known as the Nyquist theorem). If one samples at a lower rate, the encoded data become distorted, an effect known as aliasing.

With the introduction of the compact disc (CD) in 1982, a wholesale shift from analogue to digital began. On a CD, digitally encoded signals are stored as patterns of dark pits in a shiny metal layer, which are scanned by a laser. The laser reffects from smooth areas but not from the pits; the CD player interprets the reffections as 1s and the non-reffections as 0s, and the strings of 1s and 0s encode the audio information as a sequence of binary numbers. Originally, CDs were claimed to have the impressive though not especially useful capability to play even if coated with marmalade. The claimants neglected to mention that the marmalade had to be applied to the upper side of the CD, as the coding is on the underside.

Nowadays of course, music is routinely bought, stored, and played without using physical media—audio files can simply be downloaded to a computer and played through a wide range of output devices. Often, the computer is part of an MP3 player (MP3 is short for 'Moving Picture Experts Group-1 or Moving Picture Experts Group-2 Audio Layer III').

The magic of MP3 audio files is how small they are: about one-tenth the size of equivalent CD files, which means a minute of MP3 music can be squeezed into a megabyte. This impressive reduction is achieved partly by a technique called Huflman coding, in which the symbols that appear most often are coded in the shortest possible way, and partly by coding more fully those frequency bands which people will notice most if they are disrupted (mainly speech frequencies), and providing only sketchy versions of those frequencies of less concern to us.

Because MP3 players take into account both the music and the listener in deciding what information to leave out when the song is compressed, Sterne has concluded that:

> the MP3 carries within it practical and philosophical understandings of what it means to communicate, what it means to listen or speak, how the mind's ear works, and what it means to make music. Encoded in every MP3 are whole worlds of possible and impossible sound and whole histories of sonic practices . . . . MP3 encoders build their files by calculating a moment-to-moment relationship between the changing contents of a recording and the gaps and absences of an imagined listener at the other end. The MP3 encoder works so well because it guesses that its imagined auditor is an imperfect listener, in less-than-ideal conditions. It often guesses right.

Effcient coding of music exploits the fact that, over billions of years, our hearing systems have evolved to respond to the sounds which are of most relevance to us. This, combined with limitations due to the nature of sound, restricts our immediate access to the world of sound to a frequency range which is only a thin slice of what is actually out there. Those unhearable realms are the subject of Chapter 6.

Chapter 6 Ultrasound and infrasound

Bat sounds

Many textbooks will tell you that you can hear from 20 Hz to20 kHz. Don’t believe them: if you are over twenty, you are probably deaf to sounds above 17 kHz. The high-frequency limit of our hearing declines so significantly and so predictably with age that a youth-repelling generator of higher-frequency sound called the Mosquito has been used by irate shopkeepers since 2009, as only teenage ears (and those of children and babies) are still capable of detecting the annoying tones the device produces. However, it is probably fair to say that we who cannot hear anything about17 kHz are not inconvenienced by the fact, showing that it is of little evolutionary benefit to hear such high notes, otherwise we would have developed more robust hearing systems.

Nevertheless, we are missing out—not because the ultrasonic soundscape is an especially rich one, but because we cannot exploit some handy physical properties of sounds that are negligible at those frequencies that we can hear. The bat, however, is not so handicapped.

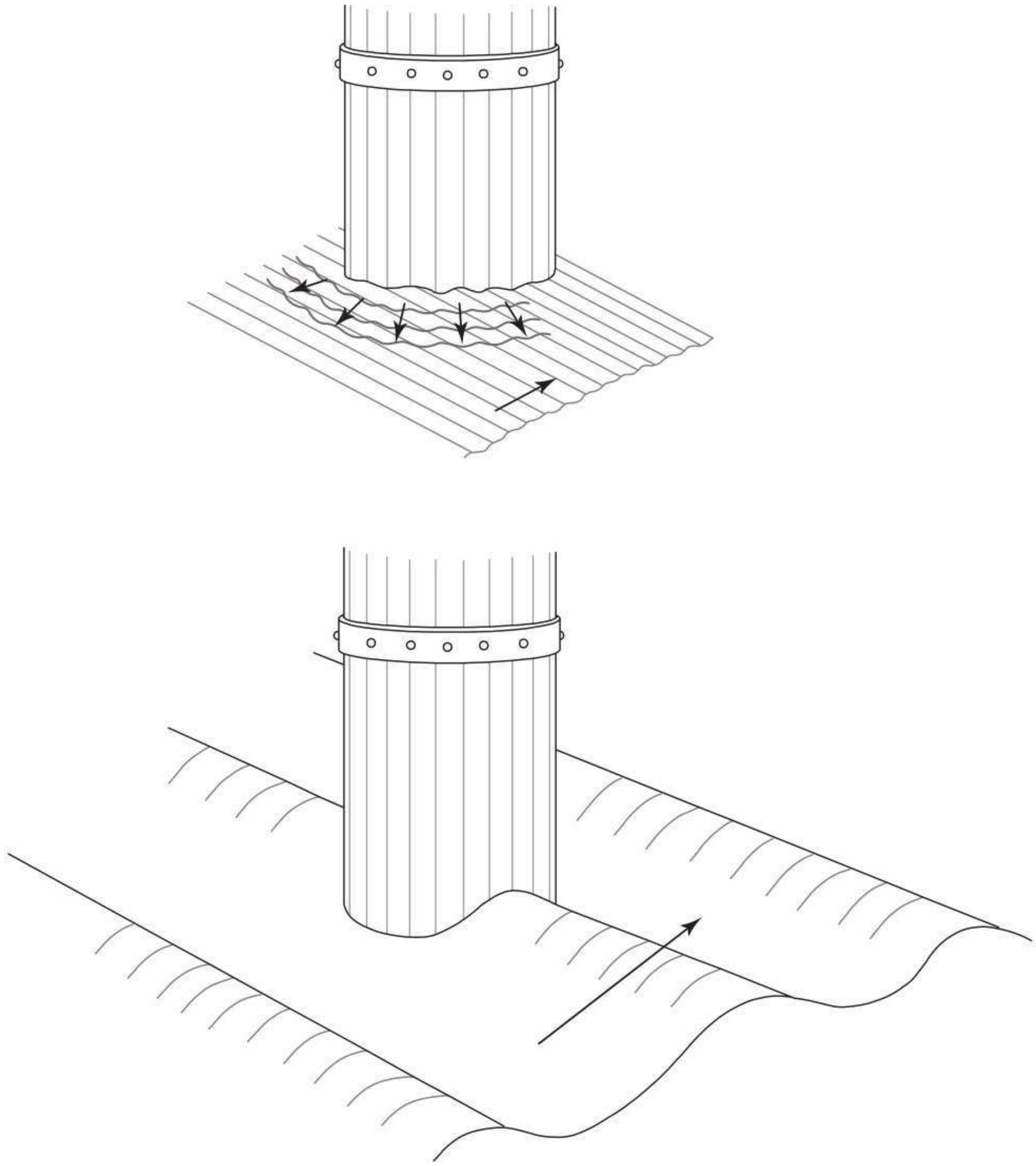
The sophistication with which bats employ ultrasound is astounding. In pitch darkness, an Egyptian fruit bat (Rousettus aegyptiacus) with an 80 cm wingspan can easily ffy between vertical rods just 53 cm apart without touching them. To achieve such feats, and to hunt, bats use echolocation—

generating ultrasounds and timing the delays until they hear the echoes—which inform them of the distances of nearby objects.

Sightless humans approximate this, judging distances to walls from echo delays, but what they end up with is not a vision of the world. It cannot be, at the wavelengths they can hear: imagine a jetty standing in a calm lake. A light breeze blows, forming ripples about a centimetre in wavelength, which bounce ofl the jetty’ s columns in circular patterns (Figure 21). Later, a gale sets in, making much longer waves, which roll past the columns as if they’ re not there. For waves of sound, the same principle applies: they are only aflected by (and hence can only detect) obstacles larger than they are.

So, if a bat tried to echo-locate a moth using an audible tone—middle C, let’ s say, which has a wavelength of 132 cm—it would only succeed if the moth were well over a metre across.To detect a 1 cm moth, sound waves of at least 33 kHz are required. In fact, bats typically produce sounds of 80 kHz (smaller species produce higher frequencies, and the full range is about 40 kHz to 120 kHz). Such high frequencies allow them to deal not only with isolated objects, but with complicated three-dimensional distributions of twigs, leaves, and insects, many in relative motion to each other—and all in relative motion to the bat.

But it would do the bat no good to generate a continuous ultrasonic whine: the result would be a bewildering montage of overlapping, interfering echoes, impossible to decode. What bats need are pulses of sound, and they must be brief: if two moths are a centimetre apart, then, to produce separate echoes, a sound pulse must have passed beyond the closer one before reaching the second, and hence must be less than a centimetre in length, and therefore less than 30 microseconds in duration.



21. Ripples vs waves.

It is impossible for bat or machine to produce a short-duration pulse with a single wavelength, unless that wavelength is much shorter than the pulse itself. Instead, bats produce very sharp clicks. Thanks to Fourier analysis, we know that such a click is equivalent to a mixture of waves

of diflerent frequencies (Figure 12), and the sharper the click (that is, the quicker it rises from silence to maximum), the wider the range of frequencies produced—and thus the more accurately distances can be judged.

So is the ideal sound for a bat an extremely short-duration click? Not really: a short click means a low-energy one, which cannot travel far without fading. Humans faced this problem themselves in designing radar and sonar systems, and cracked it in 1960 with the invention of frequency sweep. This utilizes a relatively long pulse of gradually increasing frequency. The fairly long duration allows plenty of energy and hence range, and the changing frequency means that echoes from objects at diflerent distances are distinguishable by the diflerences in their frequencies.

When bats use this frequency sweep technique, the pulses last about 2 to 3 milliseconds, during which the frequency falls (rather than rising as in the human version) by about an octave. The more such pulses a bat makes, the more information it receives, and so it modifies its click rate depending on the challenges it faces, from around ten pulses a second when cruising to as high as one hundred in a half-second period, when the surroundings become complicated or when nearing prey.

As discussed in Chapter 2, if the object from which a sound echoes is in motion (relative to the sound source), the frequency of that sound will be shifted by the Doppler

eflect. Both human and bat sonar systems exploit this to determine the velocities of such objects, with the distinction that, while our systems measure the change in frequency, bats modify the frequencies of their output pulses until the echoes they hear are the same as they would be, were the target stationary (so, for example, if a bat is approaching an object, the frequency of the echoes from that object rise, so the bat reduces the frequency of its output by an amount that brings the frequency of the echoes back down to match the bat’s original frequency).

A different kind of sound

Box 9

Diffraction of sound through an opening: (plane) angle of spread, θ, is given by $\sin(\theta/2) = \lambda/d$; diameter of the opening d, wavelength λ.

Another physical diflerence between ultrasound and the audible range is that ultrasound readily forms beams, which is of considerable advantage to bats. An 80 kHz tone passing through a 1 cm opening will form a conical beam which spreads to about 90 cm wide at a 1-metre distance (Box 9). In bat species which project their ultrasound through their nostrils, interference eflects between the two sources mean that the beams are narrowed further. Not only does this concentrate the acoustic energy, so allowing greater distances to be probed, it also reduces the number of distracting side-echoes. A 2015 study of bats approaching drinking pools showed that the bats opened their mouths wider

when they were close to the water, presumably for this latter reason.

Evolution could no doubt have furnished bats with generators and detectors that work at even higher frequencies, but the absorption of such sounds by the air is an insuperable barrier. Experimentally it has been determined that, at the sort of conditions bats love best (25° C, 50 per cent relative humidity), 100 kHz sounds are absorbed by 3 dB per metre: that is, they fall to about half of their original intensity over that distance. Conversely, 30 kHz sounds are absorbed only at the modest rate of 0.7 dB per metre, which means their intensity falls by about15 per cent. (The increase of sound absorption with frequency is the main reason why thunder sounds less crackly and more booming as distance increases.)

Box 10

Absolute gas temperature (that is, as measured in kelvins) is proportional to the kinetic energy $\frac{1}{2}mv^2$, molecular mass m, molecular velocity v.

Why does this happen? Air is composed of molecules, all randomly drifting around at a range of velocities and frequently bouncing ofl each other. On a hot day, these velocities increase—in fact, temperature is just a measure of the velocity of a large group of molecules (Box 10). A sound wave is a sequence of alternating high and low pressures moving through the air, so at any given location reached by the sound, the air molecules will first bunch up

close together and then spread apart, before bunching together again. When molecules bunch, they slow down, just as a people hurrying through a crowded station concourse in diflerent directions go slower the more of them there are. When this happens, the energy of the molecules changes form: although they move around more slowly (their kinetic energy falls, in other words), they spin and stretch more (so, their internal energy increases).

An analogy might be to pairs of cricket balls linked by strong steel springs—a fairly accurate model of a diatomic (2-atom) molecule, like the molecules of nitrogen (N_2) and oxygen (O_2) which together make up 99 per cent of our atmosphere. As soon as the compressional part of the sound wave has passed and the molecules move apart again, they spin and stretch less, and move around faster once more. Play middle C, and the sound waves produced force the energy to switch back and forth between the kinetic and internal forms 262 times a second.

But, if the frequency is sufflciently high, the time available becomes so short that a molecule cannot complete the conversion of internal energy to kinetic energy quickly enough before it is time to reverse the process. As a result, the velocity of sound falls and the sound wave dies away rapidly. The actual frequency of the sound wave at which these changes begin depends on the medium, being far higher in solids and liquids than in gases. Other properties of the medium, especially viscosity, also contribute to this eflect.

This phenomenon is very useful to us: as the ultrasound wave fades, its energy spreads through the medium, heating it up. This heating eflect has many applications, including the internal warming of body tissues to improve blood ffow or to treat damaged muscles and joints.

Medical ultrasound

Ultrasound has many other medical uses too, both in the areas of scanning (almost everyone in the developed world is now scanned before birth) and treatment (like the removal of dental tartar, where 25 kHz sound is used in conjunction with a water jet). Unlike many other medical treatments, ultrasound can be switched on and ofl instantly, often requires low-cost technology, and usually involves minimal patient preparation. The fact that ultrasound generators are relatively portable and need little ancillary equipment means that they can be used outside the hospital environment, from diathermy (deep-tissue heating) units found in many gyms to wound cauterization systems, deployed on battlefields to save the lives of soldiers who would otherwise perish from blood loss.

Ultrasound has been used for the treatment of tumours of many kinds—including otherwise inoperable brain cancers—by a technique called either HIFU (high-intensity focussed ultrasound), or HITU (high-intensity therapeutic ultrasound). As well as destroying cancerous tissues by heating them up (to about 90° C over regions about the size of a grain of

rice), ultrasonically induced bubble formation in tumours has been used to make them more susceptible to chemotherapy. The main challenge in the use of ultrasound for treatments like these, which need precise and accurate targeting, is that the paths of the ultrasound beams depend on the densities and elasticities of the intervening body tissues. So, models of body parts made from artificial tissue-mimicking materials are used to calibrate and programme the equipment.

A more straightforward use of ultrasound in medicine is lithotripsy, in which high-power pulses simply pound kidney stones in situ, smashing them to pieces small enough to pass out with urine.

Scanning with ultrasound

One of the best-known uses of ultrasound is the scanning of foetuses. Since the diaphragm of a conventional loudspeaker could not move quickly enough to produce the megahertz frequencies required, a piezoelectric projector is used instead. Gel is applied to the abdomen, so that there is no air layer to reffect or absorb the sound. The ultrasound waves bounce ofl interfaces between media with diflerent impedances, such as bone/muscle, or skin/amniotic ffuid. Hence, by measuring precisely how long it takes for echoes to return from these interfaces (taking account of the diflering sound velocities in each medium), their positions can be worked out. By moving the beam around, a detailed three-

dimensional map can be calculated, and converted to a real-time video image.

Like most loudspeakers, a piezoelectric transducer is reciprocal—supplied with an oscillating current it produces a sound wave, and, conversely, when struck by a sound wave it generates an electrical signal. So, in a foetal scanner the scanner head acts as both the source of the ultrasounds and the detector of their echoes.

With very high-frequency sound beams relatively sharp images can be made: a 1 MHz (one million hertz) signal can image details at the millimetre scale, the exact value depending on the acoustic properties of the tissue bring viewed, and many scanners now go up to 15 MHz, or 50 MHz for eye and skin scanners.

But even this is paltry compared to the eight billion hertz signals generated by an acoustic microscope, which can consequently resolve details as small as 0.03 micron. Unfortunately, such high-frequency sounds are absorbed by almost all media before they have travelled a single millimetre, and the only useable exception is liquid helium—which will boil if not kept colder than 5 kelvin (fi268° C). The elaborate cooling systems required mean that acoustic microscopes are not cheap: the reasons they are used at all are that they can probe under the surfaces of samples, and that some materials which are hard to distinguish from their surroundings visually are highly reffective to sound.

Outside the medical field, one of the commonest diagnostic applications of ultrasound is in the detection of ffaws and cracks—in railway lines, for instance. To locate these, a series of sound bursts is sent along the object to be tested. In pulse-echo mode, the transmitter and receiver are positioned together: if a ffaw is present, the pulses are reffected and their arrival times indicate the ffaw position. In transmission mode, the detector and transmitter are separated and any changes in the pulses in transit indicate the presence of inhomogeneities in the test object. This approach can also be used for the determination of mechanical stresses in solids: because the elastic moduli of materials alter when stressed, their sound velocities also change locally.

The power of ultrasonics

Though thermal eflects of high-power ultrasound have many applications, it can also deliver its energy through mechanical eflects on the medium. The snapping shrimp provides a rare natural example of the use of the shocks and stresses made by ultrasound (together with audio frequencies) to kill. The loud and sudden click produced when the creature snaps its claw includes frequencies up to 200 kHz, powerful enough to kill or stun both prey and would-be predators alike.

Whether utilized by shrimps or humans, the mechanical power of ultrasound is usually delivered through cavitation,

which is the formation and violent collapse of tiny bubbles. Almost all liquids contain such bubbles, made either of their own vapour or of air. When the pressure of a liquid falls, these bubbles grow (which is the cause of the frothing when a pressurized container of fizzy drink is opened). Since a sound wave is a sequence of low and high pressures,

it causes bubbles to swell and shrink rapidly, and at high powers and frequencies the bubbles pulsate so violently they disrupt and implode, suddenly releasing their vibrational energy in the form of heat. The temperatures involved may exceed those on the surface of the sun, and can make the liquid glow with light (a phenomenon called sonoluminescence, which snapping shrimp also initiate).

Because the energy appears only within tiny volumes, the body of liquid does not get particularly hot. But the bursts of highly concentrated energy can be used to initiate chemical changes (sonochemistry), and to clean and sterilize submerged objects such as medical instruments. Cavitation-based cleaning is most eflective in the 20 to 50 kHz range. At higher frequencies, ultrasound also causes agitation of liquids, which dominates cleaning eflects in the range 100 kHz to 1 MHz. In practice, ultrasonic cleaning baths use both eflects.

High-power ultrasound (without cavitation) is routinely used for ffuxless soldering of printed circuit boards, in which the electrically heated tip of a soldering iron is

vibrated at ultrasonic frequencies. At still higher powers, fine wires can be welded together, heated from within by the friction caused by ultrasonic vibrations. A great advantage of this is that, since the heating eflect is restricted to the insonified (sound-filled) wires, the ultrasound does not heat up nearby components. Many other materials are welded ultrasonically; the frequency chosen depends on the size of the parts to be welded, from 60 kHz for the smallest elements to 10 kHz for the largest.

High-power ultrasound can even be used to levitate small objects. Although the forces involved are feeble, in microgravity environments such as the interiors of space stations they could be used to hold delicate instruments in position during assembly, or to prevent highly reactive chemicals from coming into contact with anything.

The outer limits: phonons

At extremely high frequencies, sounds behave in ways radically different to those with which we are familiar: just like electromagnetic radiation, in which the highest frequencies behave much more like particles than waves (hence the individual clicks of a Geiger counter in registering the presence of gamma rays), the highest-frequency sounds behave like particles called phonons. Their existence was discovered indirectly: by the late 19th century, it was known that a specific amount of heat was needed to raise the temperature of a substance by one degree. The specific heat of water is

higher than that of, say, oil, which means it takes longer to boil a kettle of water than a kettle of oil.

Gases and solids have specific heats too, but there was a puzzling anomaly: increasing the temperature of a solid takes about twice as much heat energy as raising the temperature of the same amount of the same substance in its gaseous form. This means that solids (and certain liquids) must have a means of storing heat which is unavailable to gases. The mechanism is vibration: a molecule in a solid can oscillate around its equilibrium position, like a pendulum. But, unlike a pendulum, an oscillating molecule cannot slow gradually. The laws of quantum mechanics mean that it must jump from a rapid to a slow oscillation, and when it does so it transfers a phonon of vibrational energy to another molecule. The ways in which solids conduct heat and electricity can be accounted for by the behaviour of phonons.

Infrasound

As the lower frequency limit of human hearing is approached, the intensity of just-audible sounds increases: a pure tone that is just audible at 20 Hz is nearly 300 million times more powerful (all other things being equal) than a just-audible tone at 4 kHz. Such powerful low-frequency sounds are rarely encountered in air, but solid-borne versions are commonplace on construction sites, in underground stations, near motorways, and in seismically active areas, where they can easily be felt.

Low-powered airborne infrasounds, on the other hand, surround us all the time. They are even generated when we walk: the cyclic air pressure changes at our ears due to our up and down head movements constitute an infrasound wave at around1 Hz. Sea waves generate infrasounds at around 0.2 Hz. The lowest-frequency natural sounds of all come from high in the air and deep underground: both aurorae and volcanoes produce infrasounds at around 0.01 Hz.

One of the key characteristics of infrasound is that it can travel far further than audible sound, through sea, ground, or air; in air, infrasounds from thousands of kilometres away (made, for example, by erupting volcanoes) can easily be detected, though not usually by microphones. Just as we feel, rather than hear, infrasound, so we use specialized barometers to measure it. Infrasound can also be detected by means of the temperature changes it causes.

There is evidence that infrasound has a range of eflects on humans not shared by other types of sound, including enhancement of emotional responses: when a concert of modern classical music was accompanied by infrasound, there were far more people who hated the music, and far more who loved it, than at a concert of the same music with no infrasonic accompaniment. Other experiments have shown that drivers exposed to infrasound can get very tired very quickly, and it has even been suggested as a cause of supposed hauntings, partly through direct emotional effects and partly through vibration of the eyeball, giving rise to visual disturbances.

Infrasound has proved to be a reliable means of detecting bolides, which are meteoroids that explode in ffight. Since they travel supersonically through the atmosphere, they generate infrasound-rich sonic booms as they fall. More infrasound is generated when the bolides explode. All these airborne infrasounds generate solid-borne versions when they reach the Earth, and surface waves are made when the bolide fragments crash to Earth. The frequencies, timing, and amplitudes of all these sounds can be combined to give a detailed analysis of the path, motion, and energy of the bolide.

However, infrasound is usually of more consequence to us when it travels through the Earth—and ultrasound really comes into its own underwater. It is to these media that we turn in Chapter 7.

Chapter 7 Sound underwater and underground

Second sight

If our world had always been blanketed in fog, we would probably make little use of sight, and rely on our ears instead. If the fog proved opaque to radio waves, our telecommunication systems would no doubt have taken an acoustic turn, too. The ocean is a world of just this kind: visibility is poor due to the many suspended particles in the water, radio waves quickly die. And the water is full of sound: fish, marine mammals, human divers, submarines, and underwater robots all use it for subtle and complex forms of sensation, communication, and sometimes attack.

All this would be news to a scientist living only a century ago. It’s true that Leonardo da Vinci pointed out around 1490 that ‘If you cause your ship to stop, and place the head of a long tube in the water, and place the other extremity to your ear, you will hear ships at a great distance from you’, but no one took much notice of him.

It was not until World War I that listening to the underwater sounds of vessels was attempted, and not until World War II that the richness of the underwater soundscape became clear. This happened partly through two bizarre events. In 1942, acoustic buoys in Chesapeake Bay, which had been deployed to detect German submarines, all alerted at

once. A ffotilla of destroyers enthusiastically depth-charged the area—but no tell-tale oil slicks appeared and the only victims were thousands of fish. Later that same year, all along the western coasts of the US, every acoustic mine (set to protect ports from anything with a propeller, in fear of a Japanese invasion) detonated at once—and again, the sole result was a multitude of dead marine life.

If the US military had known a little more about a fish called the croaker, much embarrassment and many dead fish could have been avoided. For all its undistinguished appearance, this small brownish creature has a very loud voice indeed—something like a magnified woodpecker—and croaker colonies give voice every dawn, all together, much like birds.

Noisy fish were an almost complete surprise to biologists. There was a hitherto unquestioned belief that the depths were as silent as they were proverbially said to be. In fact, while the oceans had been routinely exploited for coastal and international travel and as a source of food, what lay beneath the surface was largely unknown and apparently rarely even speculated upon, beyond wild tales of sea monsters and legends of sunken cities. The dark underwater world seems to have been regarded as wholly alien and no place for people—even sailors regarded the depths as forbidden territory and most of them did not even learn to swim. While artists and writers—and psychologists too—often viewed the sea as the wild domain of raw nature, their

attention was confined to its rough surface alone. Beneath that surface was a negative region, assumed to be as devoid of sound as it was of light.

However, from the early 1940s, listening with hydrophones (underwater microphones) revealed just how noisy the 'silent deep' really is, with a cacophony of sounds extending well above and below the human frequency range, from a vast multitude of sources, most then unidentifiable. The loudest source in some areas turned out to be colonies of snapping shrimp. In fact, the upper layer of the sea is never quiet, with a background generated by breaking waves, rain, and lightning as well as by marine life. In shallow areas, swirling sediment is an additional, omnipresent source of sound.

Hearing underwater

There are two reasons for the apparent silence of the sea: one physical, the other biological. The physical one is the impedance mismatch between air and water, in consequence of which the surface acts as an acoustic mirror, reffecting back almost all sound from below, so that land-dwellers hear no more than the breaking of the waves.

Submerge your head, and the biological reason manifests itself: underwater, the eardrum has water on one side and air on the other, and so impedance mismatching once more prevents most sound from entering. If we had no eardrums (nor air-

filled middle ears) we would probably hear very well underwater.

Underwater animals don’t need such complicated ears as ours: since the water around them is a similar density to their ffesh, sound enters and passes through their whole bodies easily, without the need of pinnae to coax it in, nor of drums or windows to transfer it from one medium to another. Fish do have ear bones, called otoliths. They are made of calcium carbonate, the high density of which provides a suflcient impedance difflerence to allow vibration by sound waves. This motion is transferred to stereocilia growing on hair cells, which, like ours, send nerve signals to the brain. Other patches of hair cells, called neuromats, are distributed on the skins of fish.

Two other structures augment the hearing of some fish. The first, the swim bladder, is an air-filled sac which functions like the ballast tanks of a submarine, changing the buoyancy of its owner as required, so that the fish can sink or rise without muscular effort. The swim bladder vibrates readily to sound waves, and functions as a fairly sensitive hearing organ up to about 3 kHz. It has, however, a great disadvantage: being a single, symmetrical organ, the swim bladder provides no information about the direction from which sounds arrive.

It is supplemented by a second structure: the lateral line, a fluid-filled tube that runs along the sides of the

fish and functions as a direction-sensitive sound detector at low frequencies (~160 Hz to 200 Hz). Unlike the stereocilia in our ears, which wobble when sound waves induce vibrations in the basilar membrane to which they are attached, lateral line stereocilia are pushed and pulled directly by incoming sound waves, which means that the direction of the sounds are sensed directly (and also means that fish sense the motion of water molecules, rather than the pressure changes that we do). This makes fish very hard to creep up on.

Techniques and transducers

The earliest serious technological use of sound underwater was the sounding bell system, in which underwater bells placed near ports could be detected by ships fitted with primitive hydrophones in the form of carbon microphones in waterproof casings. A watchman listened on the ship in stereo and could guide the ship to port even in poor visibility. This system was fitted to many ships in the period 1875- 1930, including the Titanic and Lusitania.By 1923, there were thirty bells around the UK coast. From the 1910s, the system was gradually replaced by echo sounding, which involves making a sound underwater, timing its echo, and working out the distance from a knowledge of the velocity of underwater sound: another example of the pulse-echo technique.

From echo sounding, modern sonar (sound navigation and ranging) systems developed. In an active sonar system, short-

duration sound impulses (‘pings’) are projected from a vessel to echo from objects in their path. As well as determining distance, changes in the frequency of the received pulses are used to calculate the relative velocity of the source (using the Doppler efflect).

Passive sonar systems simply listen for underwater sounds, especially those of shipping. Automatic recognition techniques allow difflerent types of shipping to be identified from their engine sounds or even from the humming of their electrical systems: each vessel, in fact, has its unique signature or acoustic fingerprint. This technique became very important in the Cold War for recognizing and tracking enemy ships and submarines. Active and passive sonar systems are sometimes deployed on buoys (sonobuoys), which are equipped with radio systems to report their findings.

The hydrophone is the key instrument for underwater acoustic work; almost all those in use today are piezoelectric, usually based on a synthetic ceramic called PZT (lead and titanium zirconate). Unlike microphones, hydrophones must sometimes be very large, in order to achieve directionality at low frequencies. For this reason, the sides of some submarines are almost completely plated with hydrophones. These ffank arrays, as they are called, are usually made of polyvinylidene diffuoride (PVDF).

The underwater equivalent of the loudspeaker is known as a projector. Projectors have a limitation that does not apply

to loudspeakers: when the surface moves inwards to create the rarefaction phase of a sound wave, cavitation results if the rarefactional pressure is lower than that of the surrounding water. The bubbles scatter and absorb the sound, silencing the projector. The greater the depth, the higher the water pressure becomes, so the more sound power the projector can produce before the onset of cavitation.

For high-power applications, such as geophysical surveying for oil and gas exploration, airguns are used to produce underwater sound. In these, a small cavity is charged with compressed air and a relay releases the pressure suddenly, rapidly forming a large air bubble and an accompanying burst of sound. The pulse frequency is around 20 Hz to 200 Hz and the amplitude is very high (probably the ‘loudest’ man-made source in the ocean excluding large explosions). The sound travels through the seabed and reflects from the interfaces between the rock layers below. Very long (up to 10 km) arrays of hydrophones are towed near the surface to image the reflected sound, and the results are processed by computers to provide a three-dimensional map.

Though projectors are in principle just hydrophones in reverse, their physical design is often difflerent. One of the most widely used types is the Tonpilz (German for ‘singing mushroom’) transducer, in which several piezoelectric PZT discs are sandwiched between electrodes to make a stack which is terminated by a conical or cylindrical headmass, the business end of the projector. Tonpilz

transducers can generate frequencies in the 2 kHz to 50 kHz range.

For many underwater applications, including signalling and distance sensing, directional sounds are required. Just as in air, a sound source will be naturally directional if the sound waves it produces have smaller wavelengths than the transducer face. But because the velocity of sound is about five times greater in water than in air, the wavelength corresponding to a particular frequency is also about five times greater than its airborne equivalent, so directionality is harder to come by.

An elegant way to make a directional sound source is the parametric array. If two sound sources generate waves which diffler just a little in frequency, then waves with that difflerence frequency will be produced, along with others whose frequency is the sum of those of the sources. The wavelength of the difflerence wave can be as long as required, but it maintains the directionality of its parent waves.

Parametric arrays exploit the fact that sound velocity depends on density. At high sound powers, the pressure in the compressions becomes very large, increasing density significantly and therefore brieffy speeding up the sound wave; the reverse happens in the rarefactions. The efflect of these velocity changes is to distort the wave from its usual sinusoidal form.

This is a common efflect in high-power ultrasound, too. As Auguste Fourier showed, a non-sinusoidal wave is equivalent to a sum of component sinusoids. In the case of the parametric array, these components include the original waves, together with the sum and difflerence waves: the difflerence wave being the one of interest. Parametric arrays can also be used in air, to make audio-frequency sound more directional.

Although there is little that electromagnetic radiation does above water that sound cannot do below it, sound has one unavoidable disadvantage: its velocity in water is much lower than that of electromagnetic radiation in air, which means that scanning takes far longer. Also, when waves are used to send data, the rate of that data transmission is directly proportional to the wave frequency—and audio sound waves are around 1,000 times lower in frequency than radio waves. For this reason ultrasound is used instead, since its frequencies can match those of radio waves. Another advantage is that it is easier to produce directional beams at ultrasonic frequencies to send the signal in only the direction you want. However, a disadvantage is that absorption increases with sound frequency, so the range is limited.

Sounds worldwide

The distances over which sound can travel underwater are amazing. It is believed that before the proliferation of engine-powered vessels, Antarctic whales could be heard by

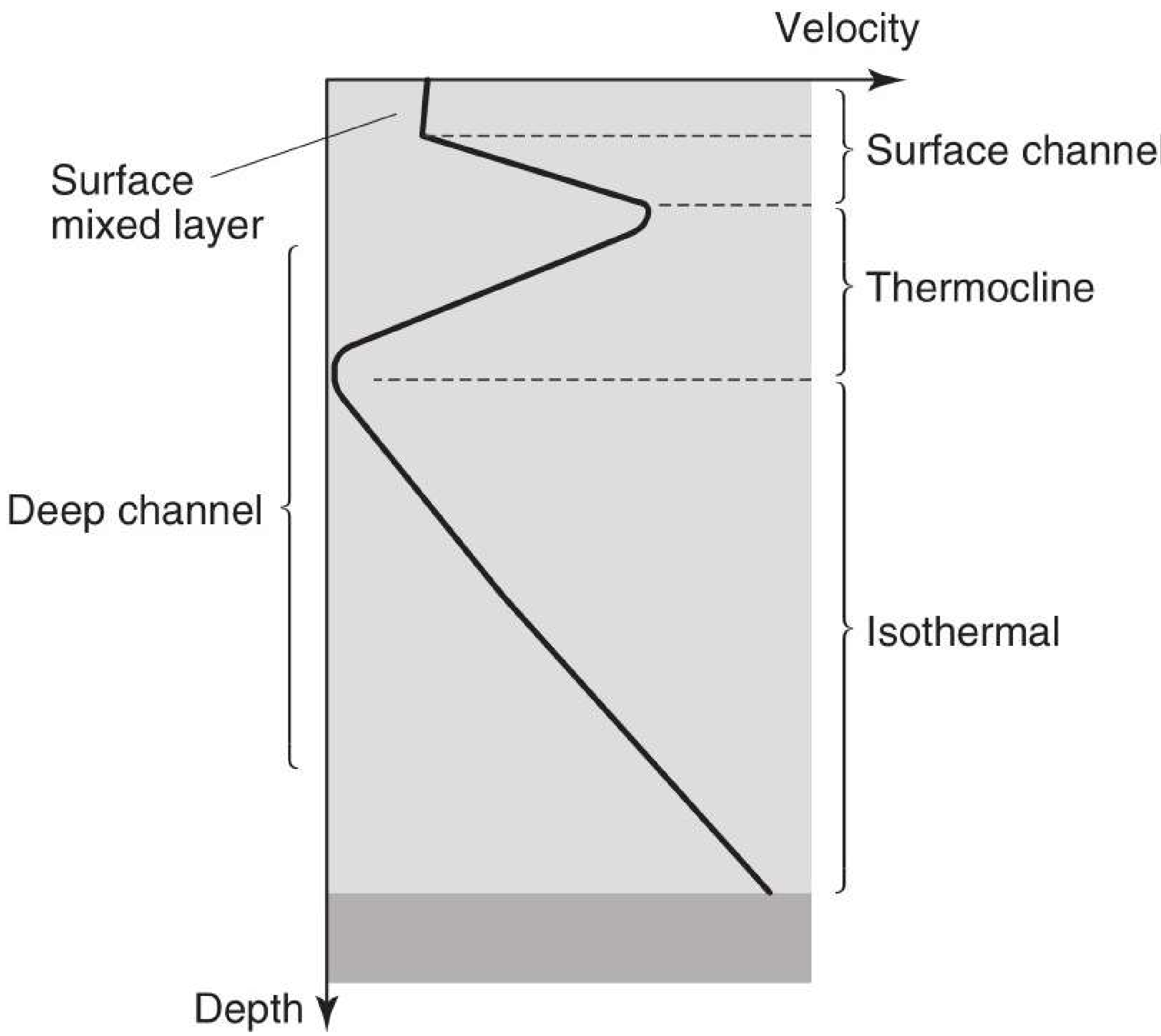
their Arctic cousins. Such vast ranges are possible partly because sound waves are absorbed far less in water than in air. At 1 kHz, absorption is about 5dB/km in air (at 30 per cent humidity) but only 0.06 dB/km in seawater. Also, underwater sound waves are much more confined; a noise made in mid-air spreads in all directions, but in the sea the bed and the surface limit vertical spreading.

Box 11

Sound velocity in water: the most accurate (to 0.2 ms^{1}) equation is the Leroy/Robinson formula: $c = 1402.5 + 5T - 0.0544T^2 + 0.00021T^3 + 1.33S - 0.0123ST + 0.000087ST^2 + 0.0156Z + 0.000000255Z^2 - 0.000000000073Z^3 + 0.0000012(\Phi - 45) - 0.0000000000095TZ^3 + 0.0000003T^2Z + 0.0000143SZ$; Ttemperature (° C), Ssalinity (per cent), Zdepth (metres), Φ latitude (degrees).

The range of sound velocities underwater is far larger than in air, because of the enormous variations in density, which is affected by temperature, pressure, and salinity (Box 11). Levels at which sound velocity changes rapidly with temperature are called thermoclines, and they follow the same pattern in most seas: when the weather is calm, the uppermost layer of the sea is characterized by a rapid fall in temperature, and hence sound velocity, with depth. Because calm conditions are more common in summer, this is known as the seasonal thermocline. Below this is the main thermocline, where the temperature and sound velocity continue to fall with depth, independent of season. At the base of the main thermocline (the depth of which varies greatly with latitude), a steady temperature of around 4° C is reached and hardly changes at deeper levels. In this deep isothermal layer,

pressure becomes the dominant factor in determining sound velocity—which consequently increases with depth, as shown in Figure 22.



22. Underwater sound velocity profile.

So somewhere under all oceans there is a layer at which sound velocity is low, sandwiched between regions in which it is higher. By refraction, sound waves from both above and below are diverted towards the region of minimum sound velocity, and are trapped there. This is the deep sound

channel, a thin spherical shell extending through the world’ s oceans.

Since sound waves in the deep sound channel can move only horizontally, their intensity falls in proportion only to the distance they travel, rather than to the square of the distance, as they would in air or in water at a single temperature (in other words, they spread out in circles, not spheres). Sound absorption in the deep sound channel is very low (it is strongly frequency dependent, but around 0.2 dB per km for 4 kHz waves), and sound waves in the deep channel can readily circumnavigate the Earth.

The deep sound channel was exploited to set up the SOFAR (sound fixing and ranging) system, which was initiated in 1960 by the Australia-Bermuda Sound Transmission Experiment, in which explosions were set offl near Heard Island in the Indian Ocean offl the coast of Australia. They were detected in Bermuda, at a distance of 20,000 km. A new, unexpected sound was discovered by the SOFAR researchers, and later identified as the calls of fin whales (Balaenoptera physalus), who long ago discovered the existence and properties of the deep sound channel and regularly visit it to signal their distant kin.

SOFAR opened the door to the ATOC (Acoustical Thermometry of Ocean Climate) system, which calculates global sea temperatures by measuring average sound velocity over large ranges, thus helping to quantify climate changes.

Changing weather causes changing sea conditions, which lead to a range of temporary acoustic states of the ocean, including shadow zones from which most sound is excluded, and temporary sound channels, which allow long-distance propagation. Sounds which travel down the latter, like those which travel around the deep sound layer, are highly distorted en route, due to the many changes of velocity on their journeys (due in turn to variations in temperature and salinity). In the 1990s, various strange sounds of unknown origin (the Bloop is perhaps the best known) led to a range of imaginative interpretations, including some literally monstrous ones. However, the most likely source is the greatly modified sounds of distant icebergs calving.

Earth sounds

Sound waves travel readily in solids (see Box 12), as do pressure waves of other kinds. But not all seismic waves are sound waves. P(for primary) waves are longitudinal: they are sequences of compressions and rarefactions whose velocity is determined by the density and elasticity of the ground, and therefore are sound waves.

However, S (secondary) waves, being transverse, are not sound waves. Both P and S waves are body waves—they travel through the Earth, and their refraction by the layers underground provide us with information about our planet’s structure. There are also a variety of surface seismic waves, but none are sound waves.

Many large animals make and hear low-frequency sounds. The sounds make by African elephants are low simply because their vocal cords are so large and hence relatively slow-moving—so much so that some of their calls are infrasonic. This is of considerable advantage to them since infrasound travels far with little attenuation (in air, very roughly, a 10 Hz signal travels one hundred times further than a 100 Hz one, and 10,000 times further than one at 1,000 Hz).

Underground, attenuation is highly variable, but usually much lower than in air: female elephants use infrasounds to attract males (over 3 km away by air or over 10 km underground) and to contact their young. Elephants also use infrasound to detect thunderstorms (a useful water source) over 500 km away. In 2004, Sri Lankan elephants ffed the coast, probably because they detected the infrasounds of the oncoming tsunami. Elephants generate their signals either by making rumbling sounds or by stamping their feet, which also detect ground-borne sound using vibration receptors called Pacinian corpuscles.

As far as humans are concerned, the use of underground sound goes back at least to 132 ce, when the first seismic detector was made in China. It was a brass vessel with a ring of metal frogs attached, each holding a metal sphere in its mouth. The presence and bearing of an earthquake was indicated by the fall of the sphere(s) in its direction.

Box 12

Attenuation of sound in gas or liquid media is described by Stokes' Law: $\alpha = 2\eta f^2/3\rho v^3$; amplitude α ff, dynamic viscosity η (nu), frequency f, density ρ (rho), velocity of sound v.

Underground sound has been a source of fear and wonder since ancient times, reactions that evolution may have hardwired into our brains, from the experiences of our ancestors that such sounds accompany avalanches, volcanic eruptions, earthquakes, and other overwhelming natural disasters. The fact that they can be felt as well as heard adds to their impact: to feel the solid earth shift and tremble beneath one's feet is an unnerving experience. No wonder that the dark world of the underground was regarded in many cultures as the abode of the dead.

Despite its unwholesome reputation however, underground sound has long been used in war: during classical times, tunnelling enemies were sometimes detected by the sounds they made through the earth, and there is even a record of shields being clashed on to the ground and the presence of tunnels below being judged by the sounds heard. In World War I, detections from difflerent points along trench walls allowed the locations of enemy troops in their own trenches to be estimated, by triangulation. In World War II, the Polish resistance spent significant periods underground, listening for and tracking German soldiers overhead—who were meanwhile tracking them by the same means.

Underground sound detectors are called geophones. Until the late 20th century, geophones worked on the same principle

as dynamic microphones (Chapter 5): a magnet was set in motion by ground waves and induced electrical signals in a wire coiled around it. Today, MEMS are used instead: a microscopic piece of silicon is mounted in a delicate holder and vibrations due to underground sound cause it to start to move. A feedback system halts this motion and the force it applies to do so provides precise information about the sound. However, MEMS devices are relatively insensitive and are used mainly for monitoring actively seismic regions. All practical geophones are highly directional and are usually deployed to respond to sounds coming directly from below.

Prospecting

One well-established use of underground sound is prospecting. Usually an explosive charge is set offl underground (there is no underground equivalent to the loudspeaker), and the waves it makes reffect offl the interfaces between layers of difflerent materials and are detected by an array of geophones.

The measurement and localization of strain-produced ultrasounds in solids, known as acoustic emission, is used to detect the onset of fractures in all sorts of structures, from planes in ffight to the London Eye. This is known as structural health monitoring. Alerts can be triggered at the instant that even an invisibly small crack begins to form. Acoustic emission is also used to study the formation of cracks during welding processes, to detect the onset of

lesions in pipes that carry high-pressure ffuids and to determine the amount of corrosion inside reinforced concrete.

Underground nuclear tests, diflcult to identify by other means, generate distinct sounds which are mostly infrasonic. The CTBTO (Comprehensive Nuclear-Test-Ban Treaty Organization) constantly monitors the Earth using geophones (together with hydrophones in the deep sound channel to detect underwater sounds originating from the sub-ocean rocks). This network of underground and underwater detectors allows the CTBTO to police international test ban treaties, and to determine the locations of any tests that do take place.

This chapter has focussed on those areas of sound that we can't hear, even though many scientists wish we could—in Chapter 8 we turn to sounds that we cannot help hearing, even though we'd rather not.

Chapter 8 Sound out of place

The nature of noise

What is noise? This is one of those questions like ‘What is time?’, to which one might answer: ‘I know exactly what it is until someone asks me’. But it’s worse than that, for ‘noise’ means two quite difierent things. For a scientist, noise is extraneous acoustic or electromagnetic energy. When used for communication, noise is whatever is not the signal, in other words does not carry information (whether the information is a voice down a wire, or a structural element in an ultrasonically scanned foetus). The larger the amount of noise, the more diffcult it becomes to unweave the signal from it—hence the concept of signal to noise ratio.

Once such noise has been identifled, it can be fairly easy to remove. Noise in a circuit that occurs at particular frequencies (a 50 Hz hum from a mains electricity supply, for example) can often be dealt with by fllters, so that only frequencies of interest remain. Unfortunately, in electronics, a very common form of noise is white noise, a signal that fllls all frequencies of interest and sounds like a hiss when heard through a loudspeaker. This cannot be fltered out so easily, but one may instead fllter all but the frequency bands that carry the signal; there will still be white noise in those bands, but the overall signal to noise ratio will improve.

For most of us most of the time though, ‘noise’ is any sound which is unwanted by the person who is exposed to it. A trumpeter may be very pleased with his output and certainly not think of it as noise, but a neighbour might well disagree, however technically excellent it may be. And both are right—one person’ s music is another’ s noise. This disagreement goes to the very heart of why noise is so diffcult to quash. But fortunately, while almost every sound has at least a few people who love it, and there are those to whom almost anything but silence is anathema, it is also true that sounds which many of us think of as noise do tend to share three characteristics: suddenness, loudness, and tunelessness. To flnd out why this is so, we need to consider why we can hear anything at all.

Ancestral legacies

We hear because our ancestors did, and they heard because their deaf relatives died. Among the many sounds with which they would have been surrounded, a few were matters of life and death. A sudden roar, the twang of a bow string, the snap of twigs underfoot, the crack of thunder overhead were all danger signs, and what they have in common is suddenness—hence today our instant reactions to sudden noise. Sudden silence, like the hush of bird song when a predator is sighted, meant as much as sudden sound and hence causes the familiar deafening silence when a clock stops ticking, an air conditioner switches ofi, or persistent rain suddenly ceases. In fact, any unexpected change to the sounds around us can be

an annoying noise: every driver knows the irritation and concern caused by a new and unidentiflable engine noise, which an unsympathetic passenger may not be able to pick out at all, however well the driver imitates it.

But why are we annoyed by those sounds which seem unrelated to any ancestral sound that might signal danger? People talking in the street outside at midnight? MP3 players you can just hear across a train carriage?

The answer lies in a key social function of sound: to demonstrate power, in particular by claiming ownership of space. We are used to the idea that each of us has a personal bubble of space that we carry round, to which the uninvited are unwelcome. Someone who invades that space with their sounds can be as annoying as someone intruding into it physically, and the main point in both examples is intent. How much more annoying is the person sitting next to you if they phone someone to talk nonsense than if they receive a call of real concern?

Whether another person is regarded as noisy or not depends too on the relationship between them and the listener. Historians Shane and Graham White, for instance, in their studies of slavery in the US, found that the sermons of African American evangelist preachers, while appreciated by congregations, were referred to simply as noisy by white American Christians who were concerned because the sounds, and perhaps their messages too, spread far beyond the

conflnes of their venues. Cultural norms are important too: acceptable sound levels in libraries, during mealtimes, and at funerals vary greatly in difierent countries.

The sea of other people' s noise in which one is often submerged on public transport, in public spaces, or open-plan offces can be a disturbing environment. The individual sounds can be neither identifled nor located, but the hearing system constantly tries to do both, making them as hard to ignore as to escape. Even when one does manage to ignore such sounds on a conscious level, the mental processing continues, as does the stress that sometimes results—and the high blood pressure that this stress can cause.

Often, the only way to reclaim one' s private sound bubble is to flll it with a sound fleld of one' s own making, by use of an MP3 player or smartphone. According to Michael Bull, Professor of Media and Film at the University of Sussex (UK): 'iPod use can usefully be interpreted as a form of pleasurable toxicity within which the "total mediated" world of users lies a dream of unmediated experience—of direct access to the world and one' s emotions'.

There are some sounds which cannot be ignored, even though they do not really exist. These so-called earworms are usually brief snatches of music, often banal ones, which insist on repeating themselves in the mind. Advertising jingles are especially good at worming their way in. Though so widespread and seemingly simple, earworms are actually

very hard to explain. Neurologist Oliver Sacks suggests that the unusual way music is remembered may be signiflcant. Recalling a scene or event involves reconstructing it, which in turn means that the result may be very difierent in difierent ‘replays’. In recalling a piece of music, however, something much closer to a direct copy of the original is preserved, an ‘almost defenceless engraving of music on the brain’,

as Sacks puts it, over which we have relatively little control. Perhaps the fact that tunes, and hence this ‘engraving’ process, arrived so late in the evolution of the mind is a reason why such memories are hard to control.

Music can have even more profound efiects on the mind—in some people it can cause epileptic seizures, while in others it can have a signiflcantly calming efiect, and can even alleviate pain and high blood pressure. It can be especially efiective in the treatment of some mental problems, and has been used in this way since World War II. Music has also been shown to be of great relief to some who sufier from Parkinson’s disease.

Deafening

The problem—indeed the tragedy—of noise-induced hearing loss is that it may not happen until years after the damage has been done. Also, while we react immediately to dangerously high temperatures or intense light, our defensive reactions to noise are far poorer.

Our eyes are equipped with a range of protective adaptations, most obviously eye lids and the contracting pupil (or expanding iris). Why does the ear have neither? The reason for the lack of ear lids is that while deafness may shorten an animal’s life, it won’t shorten it anything like as much as being successfully crept up on by a hungry tiger. And we do have an equivalent to the pupil, though it’s not particularly efiective: the acoustic refflex is provided by two muscles in each ear called the stapedius and the tensor tympani. When a loud sound arrives, the stapedius pulls the ossicle called the stapes (‘stirrup’, named from its shape) away from the oval window, while the tensor tympani pulls on the malleus (hammer) ossicle, which is attached to the eardrum and hence stifiens the latter. As a result, sounds are mufied. The stapedius usually tenses when we speak, to stop us annoying ourselves with our own voices; the tensor tympanum does the same job when we eat, to suppress our own chewing noises.

The acoustic refflex is one cause of temporary threshold shift (TTS), in which sounds which are usually quiet become inaudible. Unfortunately, the time the refflex takes to work (called its latency) is usually around 45 milliseconds, which is far longer than it takes an impulse sound, like a gunshot or explosion, to do considerable damage. The refflex is only one contributor to TTS; the other mechanisms are still unknown.

Where the overburdened ear difiers from other abused measuring instruments (biological and technological) is that it is not only the SPL of noise that matters: energy counts too. A noise at a level which would cause no more than irritation if listened to for a second can lead to signiflcant hearing loss if it continues for an hour. The amount of TTS is proportional to the logarithm of the time for which the noise has been present—that is, doubling the exposure time more than doubles the amount.

However, very loud impulse sounds, like gunshots, will also cause instant damage, even if the total energy is low. (Reputedly, a trained audiologist can identify the permanent damage caused by a single shot flred by a patient with no ear protection.) Recovery times increase with the amount of shift; a TTS of 40 dB can take weeks to recover from. The amount of TTS reduces considerably if there is a pause in the noise, so if exposure to noise for long periods is unavoidable (at a football match, say), there is very signiflcant beneflt in removing oneself from the noisy area, if only for flfteen minutes.

It is on the evidence of many careful experiments on the efiects of noise on hearing that noise regulations, now statutory in workplaces and elsewhere in many countries, have been deflned. They place limits both on peak noise levels and durations of exposure to specifled lower levels.

One problem with identifying how many people have noiseinduced hearing loss is that everyone’ s ability to hear highfrequency sounds declines with age: newborn, we can hear up to 20 kHz, by the age of about forty this has fallen to around 16 kHz,

and to 10 kHz by age sixty. Aged eighty, most of us are deaf to sounds above 8 kHz. The efiect is called presbyacusis (literally ‘old man hearing’). Since noise-induced hearing loss usually also impacts higher frequencies and, being cumulative, becomes more common with age, we don’ t really know how much loss is avoidable.

If there has been a downside to the often laudable progress made in tackling dangerously high levels of noise, it has been the relative neglect of noise which is ‘only’ loud enough to annoy people, despite their vast numbers: the 2008 Ipsos MORI (UK) National Noise Survey found that 26 per cent of respondents were annoyed by neighbour noise, for example. Those afiected lose sleep, they lose concentration, they lose patience—they lose the joy from their lives (39 per cent said their quality of life was adversely afiected). Such noise also exacerbates mental health problems, community tensions, and social isolation.

There are few who would deny that the best approach would be to remove noise from the environment in the flrst place. How this can be done, and how efiectively, depends crucially on the source. To most inhabitants of the developed world and

many elsewhere, the following sources are likely to be the main ones:

- Air traffc
- Industrial
- Neighbour
- Neighbourhood
- Rail
- Road traffc

In addition, shipping noise and wind-farm noise afiect some areas signiflcantly, while most areas are blighted by construction noise from time to time.

Of these groups, neighbour and neighbourhood noise are particularly diffcult to deal with, partly because the maker does not class them as noise but the listener does. This diffculty has led sound historian Karin Bijsterveld to conclude that noise abatement is in the grip of a ‘paradox of control’, in which the promises of experts and politicians to control noise actually amount to passing the buck to neighbours to do so, or developing such complex formulae for limiting other forms of noise, such as that from aircraft, that few can understand, engage with, or critique them.

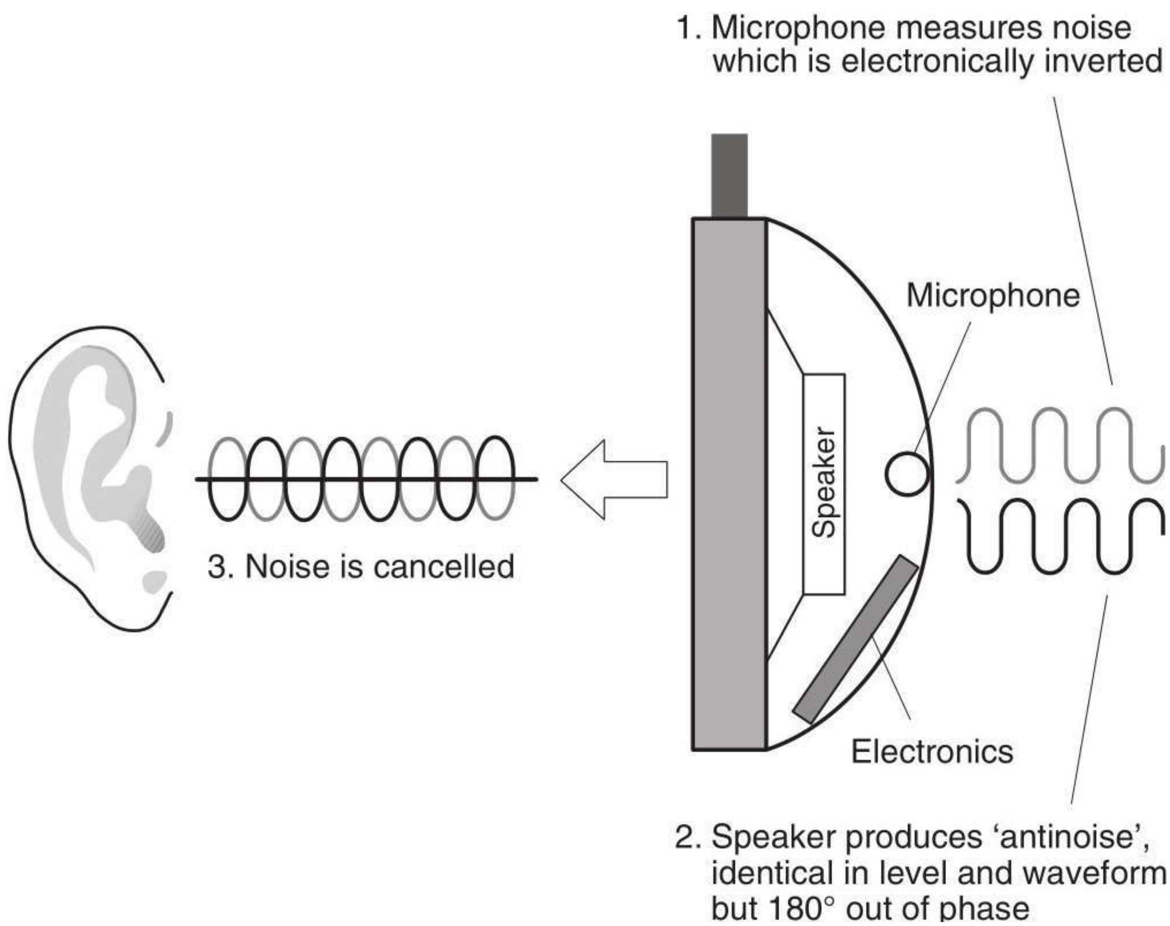
Fighting back

With the exception of some neighbour noise and some warning signals, no one would be likely to complain if noise were to be eliminated. Can this be done?

In those (unfortunately rare) cases in which noise is highly regular, consists of one or few frequencies, and is bothersome only in a small and well-deflned location, active noise cancellation (ANC) can be highly efiective. ANC relies on the fact that sound waves consist of compressions followed by rarefactions. On a pressure plot such as Figure 1, this is clear from the fact that the line goes flrst above and then below the midpoint. If a second sound wave were superimposed on the flrst such that at every point at which there is a compression in the original wave there was an equally intense rarefaction in the new one, perfect silence would result, the acoustic energy being converted entirely to heat. However, in practice this can only be achieved over very limited areas, such as plane cockpits, the driver’s head area of some cars and—perhaps most usefully—the space inside ear defenders (Figure 23).

Many highly efiective technological solutions to noise have been developed. The problem is that all of them have been applied already. Today’s vehicles are extremely quiet compared to their mechanical powers, far quieter than their much feebler ancestors ever were, but this fact is overwhelmed by the vastly greater number of them in use

today. But it is still worth looking at how today’s machines are kept so quiet, since the same principles will almost certainly be applicable in future, especially as new materials become available.



23. Active noise cancellation.

The flrst principle of noise control is to identify the source and remove it. For perhaps the single most important noise source today, the internal combustion engine (ICE), this is impossible: the noise is made by the explosions within, which are intrinsically noisy. However, as explained

in Chapter 3, our hearing systems do not measure the total energy of a sound—they are tuned to respond most strongly to frequencies around 4 kHz, and it is usually possible to shift the output frequency of ICEs to below this. The main means are the adjustment of the timing and the rate of fuel injection, to modify the combustion of the fuel so that such high-frequency components are reduced.

Having dealt as far as possible with the noise source, the next step is to contain it. Again, ICEs are challenging since they have to be connected to the air around them, to draw in oxygen and expel exhaust gases, and air paths are noise routes. To reduce noise emissions, mufiers are the answer. (A silencer is really the same thing, but is more often so-called when used to quieten flrearms.)

There are two basic mufier principles, and most ICEs use both. Absorptive silencing is achieved simply by lining the pipe with absorber so that any waves that do not travel straight down are converted to heat. Those which do travel straight down are also reduced, since their edges are retarded by the duct lining. But this effect is only slight at low frequencies, which are better dealt with by reactive mufiers. These contain Helmholtz resonators tuned to the most troublesome frequencies. Sounds at these frequencies are thus enhanced inside the mufier, allowing their energies to be absorbed there. A few mufiers also use ANC.

Over the next few decades, it seems likely that ICEs will largely give way to electric versions. There is no technical reason why these should not be almost silent, but a vehicle which approaches silently could be deadly, so all will be provided with specially generated running and warning sounds (which might also be of value for bicycles—especially those with no bells). There is no reason for these to be unpleasant, however—even an alarm need not be alarming. To attract attention, 'shushing' sounds can do the same job with less annoyance.

Jet engines, on the other hand, are probably here to stay for the foreseeable future. In their case, the source of the noise is not the burning gases themselves, but what happens when they emerge from the engine at high speed and mix with the relatively slow-moving air outside the plane. The turbulence this causes is the source of almost all the noise. Enormous efiorts have been made to smooth the mixing, mainly by increasing the circumference of the jet by corrugating it and by cocooning the exhaust jet in air of intermediate speed. No more can be done in these directions without an excessive loss of thrust, but other approaches have been considered. One idea that has met with modest success is to mount engines above the wings. Another is a strategic change, such as the use of airships for the transport of goods where transit times are not an issue.

All but the slowest or most antiquated moving vehicles are now streamlined, which avoids the turbulences that would

produce noise as well as increasing effciency. Streamlining is achieved partly by designing the vehicle with a suitable shape so that the airstream slips past it without the sudden change of direction or speed that would result in turbulence, and partly by smooth surfacing. Nonetheless, as far as car drivers are concerned, the aerodynamic noise which streamlining is intended to deal with is still the most signiflcant at cruising and higher speeds. The principal sources are usually the wing mirrors and the A pillars (the ones which hold the sides of the windscreen in place).

For those on the pavement, the loudest noise is frequently made by the tyres, especially on concrete roads. Low-noise surfaces can do a great deal to reduce this. The best available are layers of asphalt with many air-fllled pores (about 25 per cent of volume), but such surfaces are costly and not very robust. A less efiective, but more durable and cheaper alternative, is the use of thin (~2 cm) asphalt layers.

Time was when the design of machines did not consider quietness at all, and noise management was applied only as a sticking plaster afterwards or as unreliable rules of thumb (like stretching a few strings under concert hall ffloors or furnishing theatre stages with vases). Now, machines are usually well designed and carefully engineered to be at least fairly quiet—though regular maintenance of industrial machinery and careful installation of domestic appliances are essential.

When noise can be neither avoided not contained, the next step is to keep its sources well separated from potential sufierers. One approach, used for thousands of years, is zoning: legislating for the restriction of noisy activities to particular areas, such as industrial zones, which are distant from residential districts. The flrst recorded example dates from around 700 bce, when tinsmiths, blacksmiths, carpenters, potters, and even roosters were banned from the city centre of Sybaris, a Greek colony on the Aegean coast.

Where zone separation by distance is impracticable due to historical or geographical considerations, sound barriers are the main solution: a barrier that just cuts ofi the sight of a noise source will reduce the noise level by about 5 dB, and each additional metre will provide about an extra 1.5 dB reduction. Also, the barrier must be massive enough to prevent much sound passing straight through it; about 10 kg under each square metre is usually suffcient. Since barriers largely refflect rather than absorb, refflected sounds need consideration, but otherwise design and construction are simple, results are predictable, and costs are relatively low.

Temporary fflexible versions of noise barriers, known as acoustic fencing, can be placed around construction sites and, if carefully positioned and arranged to avoid leaving gaps, can be very efiective, reducing noise levels by up to 30 dB in some cases. Unlike permanent noise barriers,

acoustic fencing achieves noise reduction primarily by absorption.

The big problem with noise barriers is their unattractiveness to the eye. This can be ofiset by introducing a screen of vegetation between barrier and observer. If space permits at least 10 metres of vegetation, the plants will help reduce the noise somewhat too. At low frequencies (around 25 Hz), this is mostly due to the extra ground absorption caused by fallen leaves, and at high frequencies (above ~1 kHz) it is caused by foliage. It is best if the leaves are half as long as the sound waves, so, for both these reasons, some deciduous content is essential, (a mix of trees and bushes) reaching as high as possible (up to about 10 metres, after which refflection from large branches reduces the efiect). On windy days, the sound of such vegetation is another bonus, as a distraction from the noise.

The addition of natural sound sources (soundscaping) is especially efiective in urban parks. In such spaces the commonest way to improve a soundscape is by adding fountains, but birdsong (real or recorded) can work well too.

Quietening homes

Of all spaces, the most precious is home, because we relax and sleep there, and because it is an extension of our personal space. Keeping noise out is, therefore, a high priority: the best way to do this is through good acoustic

design, including selection of the right materials, and, crucially, their proper installation and maintenance. This is much cheaper, more aesthetically pleasing, and more successful that retrofltting noise solutions into a pre-existing dwelling.

Whether retroflted or not, the basic approaches to home sound reduction are simple: stop noise entering, destroy what does get in, and don't add more to it yourself. There are three ways for sound to enter: via openings; by structure-borne vibration; and through walls, windows, doors, ceilings, and ffloors acting as diaphragms. In all three cases, the main point to bear in mind is that an acoustic shell is only as good as its weakest part: just as even a small hole in an otherwise watertight ship's hull renders the rest useless, so does a single open window in a double-glazed house. In fact, the situation with noise is much worse than with water due to the logarithmic response of our ears: if we seal one of two identical holes in a boat we will halve the infflow. If we close one of two identical windows into a house, the infflow will again half—but that 50 per cent reduction in acoustic intensity is only about a 2 per cent reduction in loudness.

The second way to keep noise out is double glazing, since single-glazed windows make excellent diaphragms. Structure-borne sound is a much greater challenge, though if the source is within the dwelling, the room which it (or (s)he) occupies, can be treated. But this will not be cheap and will

usually require treatment of the ceiling as well as all walls, the ffloor, and the door.

One inexpensive, adaptable, and efiective solution—though maybe not the most attractive—is the hanging of heavy velour drapes, with as many folds as possible. If something more drastic is required, it is vital to involve an expert: while an obvious solution is to thicken walls, it’s important to bear in mind that doubling thickness reduces transmission loss by only 6 dB (a sound power reduction of about three-quarters, but a loudness reduction of only about 40 per cent). This means that solid walls need to be very thick to work well.

A far better approach is the use of porous absorbers and of multi-layer constructions. In a porous absorber like glass flbre, higher-frequency sound waves are lost through multiple refflections from the many internal surfaces. Using stud-mounted panels with air gaps between them and the wall will deal with lower frequencies, through refflection of lower frequency waves at the air/solid interface (yet another example of an impedance mismatch). A well-fltted acoustically insulated door is also vital.

The fioor should not be neglected: even if there are no rooms beneath, hard ffloors are excellent both at generating noise when walked on and in transmitting that noise throughout the building. Carpet and underlay are highly efiective at high frequencies but are almost useless at lower

ones, so if you have ffloor-mounted loudspeakers or an extrovert baritone in the home, something special will be required, and again there is no real alternative to bringing in an expert.

Though rarely a problem in the home, the most signiflcant issue in the design of public spaces is reverberation. Surprisingly, though it has been a bugbear for architects, performers, and audiences alike for millennia, no way to quantify it was developed until 1898, when Wallace Clement Sabine deflned the reverberation time of a room as the period necessary for the intensity of the sound to decline to one millionth (fi60 dB) of its initial level (Box 13). Even more usefully, Sabine derived an empirically based equation that allowed this time to be calculated from the size and shape of a room—so architects could predict it at last.

Box 13

Sabine's equation: $T=0.161(V/A)$, reverberation time in seconds T, room volume V, total absorption area A. This equation is still used today except for highly absorbent rooms.

However much care, money, and expertise is spent on an individual noise problem, the results will be ineffective unless both the root cause and the context of that problem are investigated flrst: it may be pointless to replace standard doors with acoustic ones if the actual problem is that they are routinely left open, and there is little point in reducing the speed limit on a trunk road if the efiect is to send traffc down residential streets. Such holistic

approaches to noise reduction extend beyond acoustic treatments: the beneflts of fltting sound absorbing panels must be balanced against their visual impact, the introduction of artiflcial sounds to distract park users from traffc noise may annoy more people than it pleases, and moving people from open-plan offces to quieter cellular ones may reduce team working. Similarly, while well-built houses can keep noise out, this necessarily involves keeping the inhabitants in, cutting ofi the sounds of nature too, and leading to alienation. As David Hendy puts it: 'whenever we withdraw into separate soundscapes . . . we make strangers of each other'.

The wider view

Despite the great importance to us of our hearing systems, their performance is not usually a matter of life and death. This is not the case for some underwater species. The prime importance of sound to many marine animals, combined with the effciency with which sounds travel underwater, means that some marine creatures are devastated by the efiects of noise, in particular whales and dolphins. The scale of the efiects is unclear but some are subtle: since whales communicate with their calves by sound signals, the mother/child link can easily be broken by extraneous sounds and the calf separated. In other cases, the startle efiect of unfamiliar loud sounds on whales and dolphins can be so extreme that rapid surfacing occurs, leading to death from the formation of nitrogen gas bubbles in the liver and

other organs (the condition divers describe as ‘the bends’).

There are many contributors to underwater noise pollution. As well as the sounds of shipping, sonar systems, blasting, and signalling all add their load. Fortunately, there is now widespread recognition of the efiects on animals; more in fact than on divers, who routinely risk severe hearing damage (although, as the efiects are often the same as those of pressure changes in the ear, the scale of the problem is hard to establish).

The interdisciplinary and public nature of attempts to battle noise, both on land and underwater, make collaboration at all levels vital, and cannot succeed without the active support of international organizations such as the World Health Organization, national governments, local authorities, specialist organizations like the UK Noise Abatement Society, and the public. One activity which should bring all these players together is noise action and awareness weeks.

As technologies have evolved, cities have grown, and vehicular travel has increased, noise sources have proliferated and noise is now a fact of life for many, whether made deliberately and selflshly via loudspeakers or unavoidably and carelessly through engines and other hardware. Although there is much that science can do to help, it must be accompanied by both education and legislation

which ensures that noise is universally regarded as the pollutant that it is.

Ultimately, noise-making must become unacceptable, rather than just annoying. While this may seem impractical, exactly such a change has taken place in the UK regarding smoking—and this took only a few years. The change came about through a combination of new laws implemented through both national and local government, and a loud and clear message that smoking causes serious illness. Given the same approach to noise, just as radical a change might be accomplished.

Of all disciplines, sound has by far the widest variety of interested parties. Understanding it enhances the work of actors, advertisers, aerospace engineers, automotive engineers, anthropologists, architects, artists, broadcasters, builders, communication engineers, composers, designers, ecologists, educators, electronic engineers, environmental health offcers, fllm-makers, historians, marine biologists, musicians, physicians, physicists, politicians, prospectors, psychologists, seismologists, sociologists, town planners, zoologists, and many others.

Hitherto, many of these domains have been separate, but over the past few decades approaches like those of soundscaping, sound studies, and the design of ever more intimate and personalized media systems have started to bring together expertise from across these flelds. Such collaborations are often hampered by the narrow training

ofiered to most practitioners, but meanwhile they are making people more aware of the vast range and power of the subject—of the multitude of practical and emotional beneflts a knowledge of sound can lead to, and of the many new applications and insights the bringing together of its many branches may yield. Such developments are to the beneflt of everyone. For all of us, the future sounds good.

References

Chapter 1: Past sounds

Details of the sound of the early universe may be found here: 〈 http://www.astro.virginia.edu/~dmw8f/BBA_web/index_frames.html 〉 ; the history of acoustics is detailed in Dayton Clarence Miller, *Anecdotal History of the Science of Sound to the Beginning of the 20th Century* (Macmillan, 1935) and Robert T. Beyer, *Sounds of Our Times: Two Hundred Years of Acoustics* (Springer, 1999). Further historical material is in: David Hendy, *Noise: A Human History of Sound and Listening* (Profile Books, 2014). The acoustics of Epidaurus are explained at: 〈 http://www.nature.com/news/2007/070319/full/news070319-16.html〉 . The most recent edition of R. Murray Schafer's classic book on soundscapes is *Soundscape: Our Sonic Environment and the Tuning of the World* (Destiny Books, 1994). The Attali material is from Mark M. Smith (ed.) Hearing History: A Reader (University of Georgia Press, 2004). Alain Corbin's classic work (translated by Martin Thom), is *Village Bells: The Culture of the Senses in the Nineteenth-Century French Countryside* (Columbia University Press, 1998).

'simultaneously a physical environment and a way of perceiving that environment', Emily Thompson, *The Soundscape of Modernity: Architectural Acoustics and the Culture of Listening in America, 1900- 1933* (MIT Press, 2002), p. 1.

Chapter 2: The nature of sound

Most of the material here may be found in Thomas D. Rossing, Richard F. Moore, and Paul A. Wheeler, *The Science of Sound* (Addison Wesley, 2001) and Daniel R. Raichel, *The Science and Applications of Acoustics* (Springer, 2006). Most cultural references are from Jonathan Sterne, The Sound Studies Reader (Routledge, 2012).

Chapter 3: Sounds in harmony

Material from this chapter is from Neville H. Fletcher and Thomas D. Rossing, *The Physics of Musical Instruments* (Springer, 2008), Daniel J. Levitin, *This is Your Brain on Music: Understanding a Human Obsession* (Atlantic Books, 2008), John Powell, *How Music Works: A Listener's Guide to Harmony, Keys, Broken Chords, Perfect Pitch and the Secrets of a Good Tune* (Particular Books, 2010), and Eric Taylor, *The AB Guide to Music Theory* (Oxford University Press, 2013).

Chapter 4: Hearing sound

Information in this chapter is largely from David Howard and Jamie Angus, *Acoustics and Psychoacoustics* (Focal Press, 2000), Lawrence J. Raphael, Gloria J. Borden, and Katherine S. Harris, *Speech Science Primer* (Lippincot Williams & Wilkins, 2003), William Yost, *Fundamentals of Hearing: An Introduction* (BRILL, 2013), and H. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models* (Springer, 2006). The Barthes material is from Bruce R.

Smith's *The Acoustic World of Early Modern England: Attending to the O-Factor* (University of Chicago Press, 2nd edition, 1999).

'My feeling is that an entire culture…', Brandon LaBelle, *Acoustic Territories: Sound Culture and Everyday Life* (Continuum, 2010), p. xvi.

'the rich undulations of auditory material…', LaBelle, *Acoustic Territories*, p. xxi.

'Sound creates a relational geography…', LaBelle, *Acoustic Territories*, p. xxv.

The statistics on worldwide hearing loss are from the World Health Organization's 'Deafness and Hearing Loss' fact sheet (number 300), Updated March 2015. It can be found at 〈 http://www.who.int/mediacentre/factsheets/fs300/en/〉.

Chapter 5: Electronic sound

Technological details are to be found in Glen Ballou, *Electroacoustic Devices: Microphones and Loudspeakers* (Focal Press, 2009), F. Alton Everest and Ken C. Pohlmann, *Master Handbook of Acoustics* (Tab Electronics, 2009). This book, and Bruce and Marty Fries, Digital Audio Essentials (O'Reilly Media, 2005), also give practical guidance, while historical material is from Greg Milner, *Perfecting Sound Forever: The Story of Recorded Music* (Granta, 2010) and Roland Gelatt, *The Fabulous Phonograph, 1877-1977* (Cassell, 2nd edition, 1977).

Material on the effects of recordings on performance are from Robert Philip, *Performing Music in the Age of Recording* (Yale University Press, 2004) and Mark Katz, *Capturing Sound: How Technology Has Changed Music* (California University Press, 2010).

'the MP3 carries within it practical and philosophical understandings . . .', Jonathan Sterne, *MP3: The Meaning of a Format (Sign, Storage, Transmission)* (Duke University Press, 2012), p. 2.

Chapter 6: Ultrasound and infrasound

Material from this chapter is from Vivien Gibbs, *Ultrasound Physics and Technology: How, Why and When* (Churchill Livingstone, 2009), Dale Ensminger and Leonard J. Bond, *Ultrasonics: Fundamentals, Technologies, and Applications* (CRC Press, 2011), Tim Leighton, The Acoustic Bubble (Academic Press, 1994), and Gillian Sales and David Pye, *Ultrasonic Communication by Animals* (Chapman & Hall, 1974). The 2015 bat study referred to in this chapter is Pavel Kounitsky et al., 'Bats Adjust their Mouth Gape to Zoom their Biosonar Field of View', *Proceedings of the National Academy of Sciences* 2015, Vol. 112, No. 21, pp. 6724 – 9.

Chapter 7: Sound underwater and underground

Most technological and physics material is from Xavier Lurton, *An Introduction to Underwater Acoustics: Principles and Applications* (Springer Praxis Books, 2010), M. A. Ainslie, *Principles of Sonar Performance Modelling* (Springer Praxis

Publishing, 2010), and Robert J. Urick, Principles of Underwater Sound (Peninsula Publishing, 2013). Biological content is from W. J. Richardson, C. R. Greene Jr, C. I. Malme, and D. H. Thomson, Marine Mammals and Noise (Academic Press, 1995). Historical material is from Robert T. Beyer, Sounds of Our Times: Two Hundred Years of Acoustics (Springer, 1999).

Chapter 8: Sound out of place

Most material is from Mike Goldsmith, Discord: The Story of Noise (Oxford University Press, 2012). Cultural references are from Smith and Karin Bijsterveld, Mechanical Sound: Technology, Culture, and Public Problems of Noise in the Twentieth Century (MIT Press, 2008).

'iPod use can usefully be interpreted…', Michael Bull, 'iPod: A Personalized Sound World for its Consumers', Revista Comunicar, 2010, Vol. XVII, No. 34, pp. 55- 63.

Oliver Sacks' account of earworms is from his book Musicophilia (Picador, 2011), p. 47.

'whenever we withdraw into separate soundscapes . . . we make strangers of each other', James Hendy, Noise: A Human History of Sound and Listening (Profile, 2013), p. 331.

Further reading

Chapter 1: Past sounds

Robert T. Beyer, *Sounds of Our Times: Two Hundred Years of Acoustics* (Springer, 1999).

David Hendy, *Noise: A Human History of Sound and Listening* (Profile Books, 2014).

Mark M. Smith (ed.), *Hearing History: A Reader* (University of Georgia Press, 2004).

Chapter 2: The nature of sound

Trevor Cox, *Sonic Wonderland: A Scientific Odyssey of Sound* (The Bodley Head, 2014).

Trevor Pinch and Karin Bijsterveld, *The Oxford Handbook of Sound Studies* (Oxford University Press, 2013).

Daniel R. Raichel, *The Science and Applications of Acoustics* (Springer, 2006).

Thomas D. Rossing, Richard F. Moore, and Paul A. Wheeler, *The Science of Sound* (Addison Wesley, 2001).

Jonathan Sterne, *The Sound Studies Reader* (Routledge, 2012).

Chapter 3: Sounds in harmony

Daniel J. Levitin, This is Your Brain on Music: Understanding a Human Obsession (Atlantic Books, 2008).

John Powell, How Music Works: A Listener's Guide to Harmony, Keys, Broken Chords, Perfect Pitch and the Secrets of a Good Tune (Particular Books, 2010).

Chapter 4: Hearing sound

David Howard and Jamie Angus, Acoustics and Psychoacoustics (Focal Press, 2000).

William Yost, Fundamentals of Hearing: An Introduction (BRILL, 2013).

Chapter 5: Electronic sound

F. Alton Everest and Ken C. Pohlmann, Master Handbook of Acoustics (Tab Electronics, 2009).

Roland Gelatt, The Fabulous Phonograph, 1877- 1977 (Cassell, 2nd edition, 1977).

Greg Milner, Perfecting Sound Forever: The Story of Recorded Music (Granta, 2010).

Chapter 6: Ultrasound and infrasound

Dale Ensminger and Leonard J. Bond, Ultrasonics: Fundamentals, Technologies, and Applications (CRC Press, 2011).

Vivien Gibbs, *Ultrasound Physics and Technology: How, Why and When* (Churchill Livingstone, 2009).

Tim Leighton, *The Acoustic Bubble* (Academic Press, 1994).

Chapter 7: Sound underwater and underground

Xavier Lurton, *An Introduction to Underwater Acoustics: Principles and Applications* (Springer Praxis Books, 2010).

Robert J. Urick, *Principles of Underwater Sound* (Peninsula Publishing, 2013).

Chapter 8: Sound out of place

Mike Goldsmith, *Discord: The Story of Noise* (Oxford University Press, 2012).

Garrett Keizer, *The Unwanted Sound of Everything We Want* (Public Affairs, 2012).

Oliver Sacks, *Musicophilia* (Picador, 2011).

John Stewart, Francis McManus, Nigel Rodgers, Val Weedon, and Arline Bronzaft, *Why Noise Matters: A Worldwide Perspective on the Problems, Policies and Solutions* (Routledge, 2011).

牛津通识课：光学

[英]伊恩·沃姆斯利 著
叶昊扬 周应秋 译

浙江科学技术出版社

著作合同登记号 图字：11–2020–461 号
Light: A Very Short Introduction
Copyright:© Ian Walmsley 2015
Light: A Very Short Introduction was originally published in English in 2015. This translation is published by arrangement with Oxford University Press. Dook Media Group Limited is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

中文版权：©2021 读客文化股份有限公司
经授权，读客文化股份有限公司拥有本书的中文（简体）版权

图书在版编目（CIP）数据

牛津通识课．光学 /（英）伊恩·沃姆斯利（Ian Walmsley）著；叶昊扬，周应秋译．—杭州：浙江科学技术出版社，2021.4
书名原文：Light: A Very Short Introduction
ISBN 978–7–5341–9414–6

Ⅰ．①牛… Ⅱ．①伊… ②叶… ③周… Ⅲ．①科学知识–普及读物②光学–普及读物 Ⅳ．① Z228 ② O43–49

中国版本图书馆 CIP 数据核字 (2020) 第 254526 号

书　　名　牛津通识课：光学
著　　者　［英］伊恩·沃姆斯利
译　　者　叶昊扬　周应秋

出　　版　浙江科学技术出版社
地　　址　杭州市体育场路 347 号
邮政编码　310006
发　　行　读客文化股份有限公司
网　　址　www.zkpress.com
联系电话　0571–85176593
印　　刷　北京盛通印刷股份有限公司

开　　本　787mm × 1092mm　1/32
印　　张　7.5
字　　数　90 000
版　　次　2021 年 4 月第 1 版
印　　次　2021 年 4 月第 1 次印刷
书　　号　ISBN 978–7–5341–9414–6
定　　价　32.00 元

责任编辑　卢晓梅　责任校对　张　宁　责任美编　金　晖　责任印务　叶文炀

图书出现倒装、缺页等印装质量问题，本社销售部负责调换

献给凯特、亚历克斯、纳撒尼尔与克莱尔。

序言与致谢

因为有光，我们才能看到这个世界。这个显而易见的事实开启了一次漫长而引人入胜的征程。在这个征程中，有关自然世界是如何运作的非凡观点得到了发展，同样得到发展的还有建立现代科学所使用的工具和方法，从而使我们的征程保持着理论与实验齐头并进的态势。

光携带的信息关乎我们周围的一切，从遥远的恒星和星系到我们身体中的细胞，再到单个原子和分子，不一而足。光也是很多能够提高我们生活质量的技术的基础：互联网是由光驱动的；世界上最精确的时钟依赖于光；那些最为微小的物体——从单个原子到活的生物细胞——都可以通过光来进行观察与操纵；至于与光息息相关的图像和显示器，更是无处不在。光还揭示了量子世界所有的怪异之处，激发了我们对世界的想象。你可能会惊讶地发现，从人类真正理解光的本质到现在，还不足100年时间。直到现在，我们仍然试图从这种理解中获得新的见解。

这本书探讨了人们是如何从无到有地建立起对光的理解，包括光的本质以及光的用途。这是一次伟大的历程，从古代开始，就有来自世界各地的人为此作出贡献：从雅典的欧几里得(Euclid)与巴格达的海什木(Al-Hazen)构想出光线的概念；到美国洛杉矶的西奥多·梅曼[1](Ted Maiman)与日本德岛的中村修二开发了新的激光器；接着通过巴黎的奥古斯丁·菲涅耳[2](Auguste Fresnel)与伦敦的托马斯·杨(Thomas Young)对光的波动性进行的研究；英国阿伯丁的詹姆斯·克拉克·麦克斯韦(James Clerk Maxwell)和德国柏林的海因里希·赫兹(Heinrich Hertz)从这些成果中发展出了电磁场概念；并最终由德国伯尔尼大学的阿尔伯特·爱因斯坦(Albert Einstein)与英国剑桥大学

的保罗·狄拉克(Paul Dirac)博采众长，通过一个新的概念“量子场”解释了这些明显不相容的概念，即光是如何能够同时既是粒子又是波的。

对光的了解每加深一步，都会带来新的应用。例如，基于对光的折射的理解，人类发明了用于矫正视力的眼镜。在光学领域，科学发现可以很快地转化成技术得以应用。光对现代世界的影响无疑是巨大的，但是却没有得到应有的重视。因此，联合国指定2015年为国际光年(International Year of Light)，来歌颂光以及它所带来的一切。

我想要感谢亚历克斯·沃姆斯利(Alex Walmsley)和拉莎·梅农(Latha Menon)，他们为本书的初稿提出了很多有用的评论，也要感谢我的同事们解答了一些具体技术的问题，还要感谢那些为本书提供插图的人。爱因斯坦曾经说过，“凡事都应该尽可能简单，但不应过分简单”(Everything should be made as simple as possible, but no simpler)。如果在本书中出现了与这个原则相违背的地方，那一定是我没有尽力遵从爱因斯坦规劝的缘故。

[1] Ted是人们对西奥多·梅曼(Theodore Maiman)的昵称。

[2] 作者误将奥古斯丁·菲涅耳的名写作Joseph。

目录

序言与致谢
01 光是什么 What is Light?
02 光线 Light Rays
03 光波 Waves
04 光的波粒二象性 Duality
05 光物质 Light Matters
06 光、空间与时间 Light, Space, and Time
07 探索光学前沿 Lighting the Frontiers
08 量子光学 Quantum Light
09 黎明时分 Twilight
术语对照
Contents
Preface and acknowledgements
List of illustrations
Chapter 1 What is light?
Chapter 2 Light rays
Chapter 3 Waves
Chapter 4 Duality
Chapter 5 Light matters
Chapter 6 Light, space, and time
Chapter 7 Lighting the frontiers
Chapter 8 Quantum light
Chapter 9 Twilight
Further reading

01 光是什么 What is Light?

光，使我们看到周围的世界。我们通过对光的感知来收集物理世界的第一手信息，并观察它们的变化。从这个角度来说，光的这种承载与传递信息的能力，也许就是它最重要也最显著的特征。

眼见为实

视觉帮助我们找到自己在周围环境中的位置，同时也帮助我们定义外部事物，为我们构建出一幅真实的世界图景。不仅如此，视觉所激发的想象力远超视觉本身。在图1里我们可以看到乔治·里奇蒙德(George Richmond)的画作《 光的创造》，它表现了光在人类心灵中占据了非常核心的位置。英语中还有很多由光的概念衍生出来的词语，例如insight(洞察力)、illumination(既可以译作光源，也可以译作启迪)、clarity(清晰，既可以表示画面或声音的清晰，也可以表示思维的清晰)。这些词语都与人的品质或是世界的物理特性有关。在拉丁语中，有这样两个描述光的词语，一个是lux[1]，一个是lumen[2]。它们分别从物质的角度与形而上学的角度对光进行描述。本书的重点放在前者，即从物质的角度来解析光的概念。在哲学、神学、心理学、艺术和文学领域，光的物理特性与其特有的诗意交织纵横，使得光也常被拿来作比喻，表达对世界的思索，毕竟几乎所有人都感知过光。正是由于光的物理特性，以及人们通过对光的思索而产生的意识观念，使光成为科学家与哲学家们数百年来经久不衰的研究对象。

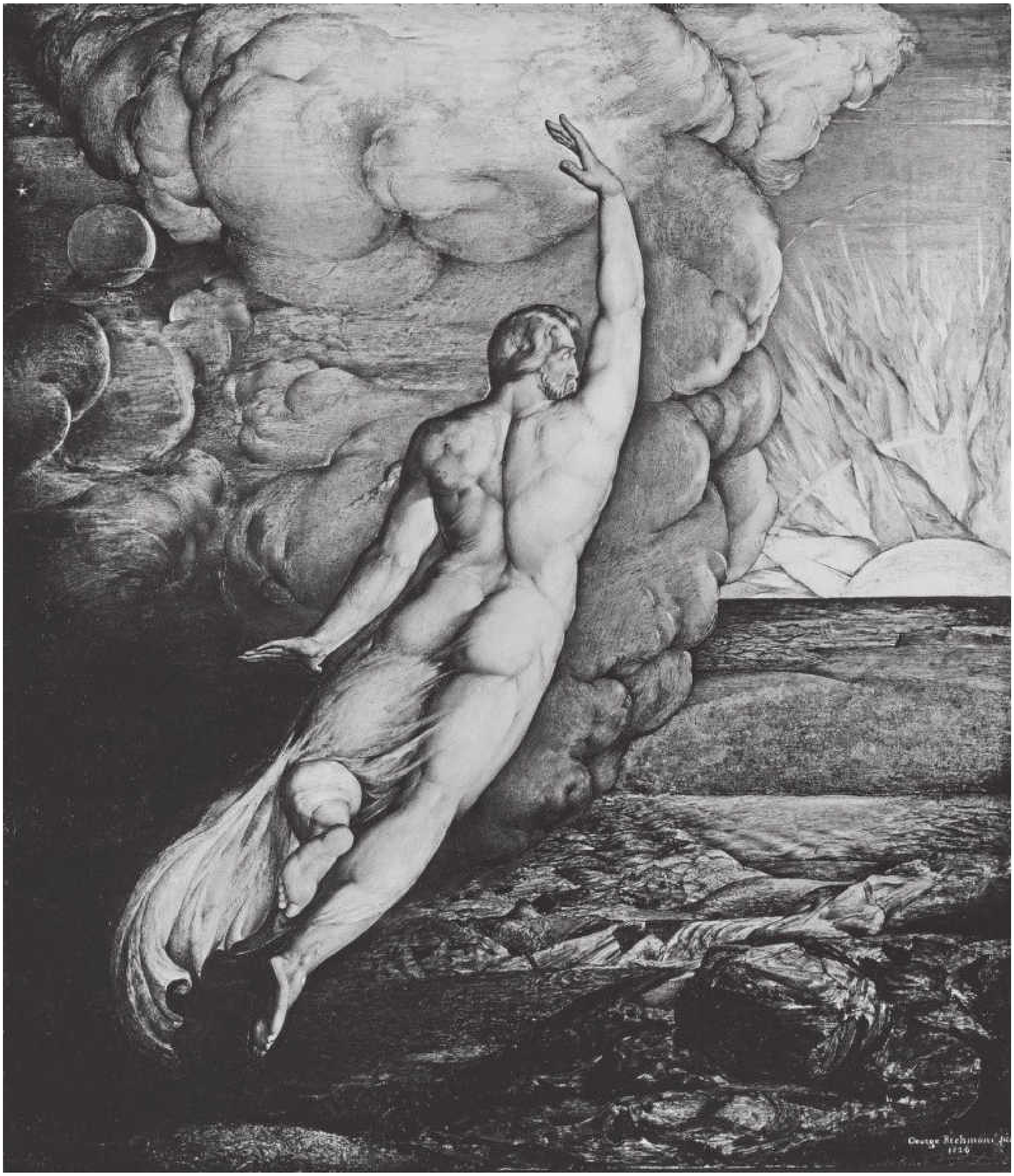


图1　乔治·里奇蒙德《 光的创造》

光赋予万物生命。毫不夸张地说，光与我们人类乃至地球得以存在的生物反应过程和化学反应过程息息相关。可以说，光塑造了我们

对周围事物的认知。从日常经验出发，我们也能理解光的这种重要性。例如，我们用光来进行照明，既可以是太阳或者月亮发出的自然光，也可以是人造光。大多数的常见光源都使用电力，但有时候我们也使用化学反应来制造光源，例如燃烧蜡烛。光源的不同特征会影响我们对周围环境的感知，因为不同的光源为它所在的物理空间营造了不同的“情绪”。

光使生命成为可能。地球上的能量主要源自太阳。光是太阳辐射能量的一种主要方式，包括可见光和我们无法直接观测到的不可见光。想象一下，你躺在阳光下的沙滩上，或者坐在室外的花园中，你身上感受到的“温暖”就是因为太阳辐射出的某种不可见光的作用。这只是光的生理效应的一个例子。

为保证生命延续，地球持续进行着一系列令人惊叹的生化反应过程。在这一过程中，光扮演着非常重要的角色，在它的助力下，“无用分子”二氧化碳被转化为“有用分子”氧气。这一反应的逆反应，即将氧气转化为二氧化碳，来自生物的呼吸以及燃料的燃烧。

在太阳光数百万年时间的作用下，地球上的生态圈以及可以提供其他能源的地质特征得以形成。没有太阳能，不管是煤炭还是石油都不可能形成。但是，我们对于这些化石能源的利用正在改变太阳光影响地球的方式。太阳辐射的一些不可见光，例如紫外线，可以被地球和大气所吸收。但是其他不可见光，例如红外线，则会被大气反射回去。同理，由地球表面产生的红外线会被困在大气层之下，导致地球表面温度升高。

利用光进行通信

自人类诞生以来，图像一直都是人类文明的一部分。图像不仅影响着我们对世界的认识，还深深影响着我们对自身所处环境的认知。

光学技术在图像方面作出了革命性的贡献。比如，通过胶片和数码摄像可以简单、快速地捕捉影像。这使得人们可以记录下人物、地点和事件来进行报道，并广泛传播出去(现在主要是通过互联网传播，而互联网的应用也离不开光学)。除此之外，它们还可以记录那些具有深远影响的事件，例如领袖与工人们的照片、令人惊叹的自然景观、毛骨悚然的战争场景，这些图像会以意想不到的方式使人们变得团结或分裂。例如：图像可以用来号召群众采取行动，激发同情心，以及使人们对共同拥有的经历进行更深刻的反思，等等。我们可以回想一下人类在月球上迈出第一步的场景(见图2)，它引发了全人类的强烈惊喜。捕获移动影像的技术为历史叙事增加了一个新的维度——你能够想象没有电视、电影和视频的生活吗？

中根本不会有所留意。我们每天都使用发光显示屏，比如电视、电

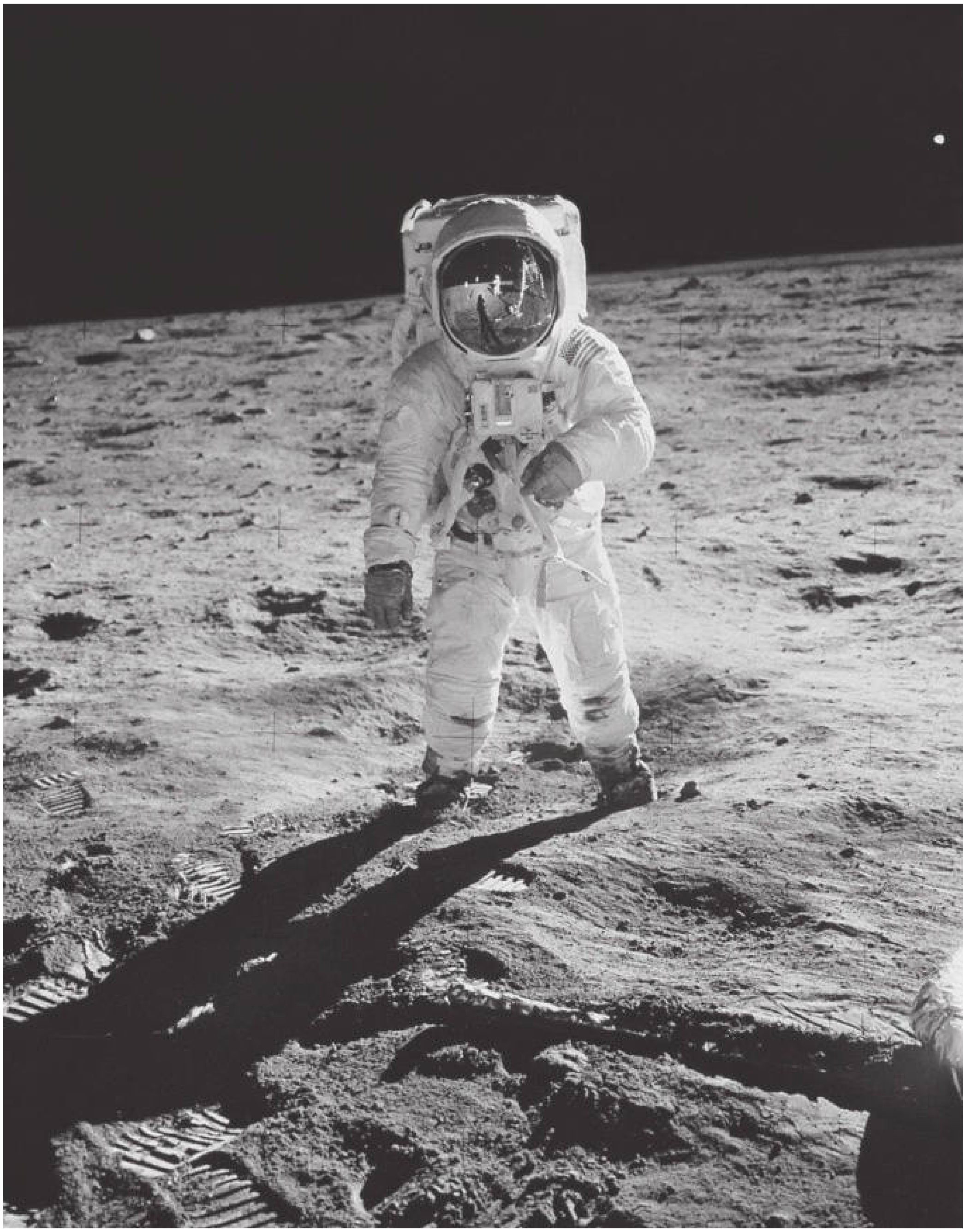


图2　尼尔·阿姆斯特朗(Neil Armstrong)在月球上为登月同伴埃德温·奥尔德林(Edwin Aldrin)拍摄的一张照片，照片中奥尔德林在月球表面行走

现在，图像的生成与传播已经如此普遍，以至于我们在日常生活中根本不会有所留意。我们每天都使用发光显示屏，比如电视、电

脑、平板和手机。这些设备都以光为媒介，为用户传送信息。几乎所有的远程通信都是以光为媒介，沿着名为光纤的细玻璃丝进行传输的。这正是互联网光纤宽带服务的基础，基于此，千家万户得以连接上互联网。即使在电视机和电脑里，光也非常重要。例如，刻在CD或DVD上的音乐、视频与图片都是通过光进行读取的。微型激光器的微型磁头被用来“读取”光盘，它将光盘上的信息转换成电信号并传输到显示屏上进行显示。我们上网、下载文件、收发邮件等活动都需要能够容纳巨大信息量的传播媒介，而唯一能满足我们要求的，就是光。

在现代社会的交通运输中，光以信号灯的形式来规范我们的交通运行。从城镇的路灯到机场的着陆灯，光都是导航系统的重要组成部分。光在车辆维护方面也很重要。例如，激光被用来对齐汽车的车轮，又比如常见的光控点火器，可以用来驱动内燃机引擎。

可以说，光以多种多样的方式，承载着创造现代生活所需要的能量与信息。

光学

所有对光进行的研究统称为光学。光学是最古老的科学学科之一，它的历史发展进程为现代科学的产生开辟了一条极其重要的道路。光学领域中提出的新想法曾经为很多不同领域研究中新观点的产生提供了灵感，比如原子和分子动力学。通过对光的深入了解，人们开发出新的技术，这些技术已经成为解开自然界中其他未解之谜的关键。例如，伽利略·伽利雷(Galileo Galilei)设计的光学望远镜不仅帮他观测到了木星的卫星，还帮助其他科学家向“太阳系中的行星均围绕太阳运动”这一新发现迈出重要的一步。而这一发现反过来又极大地促进了万有引力定律(解释行星运动的规律)的发展。

光学这一学科可以追溯到公元前4世纪，一些希腊哲学家对光进行了研究。这之后的两千多年中，光学一直持续蓬勃发展。令人惊讶的是，尽管已经有很多具备卓越才能的学者前赴后继地研究，光学在两千多年后的今天仍然不断地绽放出新的光彩。时至今日，仍然有很多新的光学现象不断被发现。在过去的20年间，有十多个诺贝尔奖被颁发给了与光相关的科学研究，可以说，光学仍然处在现代科学的前沿。例如，在超低温条件下如何在超短时间内对原子和分子的运动进行控制和测量的研究，使钟表的时间精确度提升了1000倍；又如，研究如何才能够观测到活细胞的内部，从而使我们能了解其中的变化情况。

光是什么

要讨论光是什么，可以从光非常常见的一些性质入手：亮度、强度、颜色和温度。这些可感知的特性都说明了光是一种物质实体。但是，光到底是什么呢？

我们可以用家用灯泡作为例子，先来谈谈光的亮度。家用灯泡的**功率**通常有几十瓦(功率的单位为瓦特，简称为瓦，代表每秒所消耗的能量)，具体数值根据灯泡型号不同而不同。一个50瓦的灯泡足以照亮整个房间，而汽车前照灯的功率一般略大一些，在60瓦到100瓦之间。足球场上用来照明的泛光灯的功率则更大，高达几千瓦。之后我会详细讨论光是如何由这些不同的光源产生的，但是通过功率的大小，我们已经对光的亮度有一个具体的概念了。毋庸置疑，太阳是亮度最大的光源之一，它辐射的能量巨大，功率超过了1025瓦(在数字1后面跟着25个0!)。由于太阳的亮度是如此之大，即使它距离我们非常遥远，我们仍然不能直视太阳。

以上讨论的光，与我们相距越远看起来就越暗。因此，功率并不是衡量亮度的唯一指标。在某种程度上，亮度与我们从光源那里接收

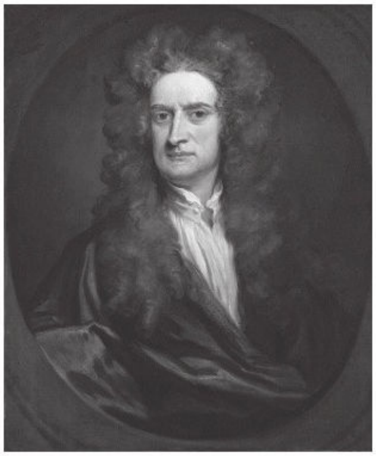
到的能量的比例有关。例如，一支激光笔的功率比灯泡的功率低很多，通常只有千分之几瓦，但是它照射在屏幕上的时候看起来非常亮。这就引出了下一个与光的亮度有关的概念。

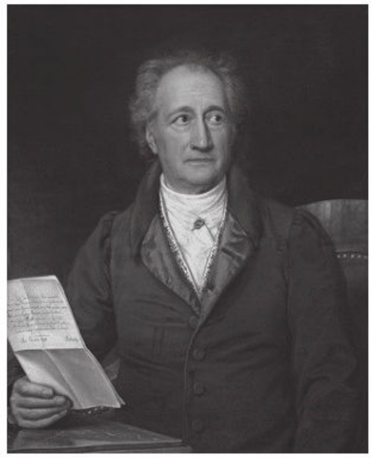
这个重要的概念就是光的**强度**(更准确的表达是“**辐照度**”，但是人们更熟悉的可能是“强度”这个术语)，即接收器每单位面积上接收到的光的能量。光的强度取决于光源的聚光能力。激光笔发射出的光看起来很亮，这是因为它的光束聚集在一个很小的点上，相比而言，太阳光则在一个很大的区域内扩散。因此，即使太阳输出的能量很大，但是它发出的光的强度却不及一支激光笔。

描述光源聚光能力的基本特征被称为“**光源相干性**”。这与光源向某个特定方向发射光的特性有关。比如，太阳和灯泡总是向各个方向辐射光，这就是为什么我们在地球的任何角落都能看到太阳，在房间的任何地方都可以看到灯泡。但是激光笔只朝一个方向发射光，即激光笔所指向的那个方向。如果激光没有照射在某个表面，人们就无法看到激光，这是因为激光束具有明确的指向性。所以激光笔是很好的相干光源，灯泡则是非相干光源。

光的另一个决定性特征，可能也是它最明显的一个特征，就是光的**颜色**。彩虹是雨水与阳光的相互作用而显现出来的多彩色带，色带由蓝色逐渐变化到红色，它体现了色谱的基本概念。事实上，彩色视觉模型的发展是光理论发展的一个中心环节。色彩不仅与感官紧密相连，还与物理学密切相关。这一点可以从艾萨克·牛顿(Isaac Newton)爵士(见图3a)与约翰·沃尔夫冈·冯·歌德(Johann Wolfgang von Goethe)(见图3b)分别做的探索颜色本质的实验中得到证实。牛顿是18世纪早期科学领域的领军人物，他在著作《光学》中对光进行了定义，这一定义在长达两个世纪的时间里受到了广泛认同。歌德则是

18世纪后期文学界的领军人物，他将科学理念融入作品中，但是他认为牛顿对于光的本质的认识大错特错。


a


b


c

图3　艾萨克·牛顿(a)、约翰·沃尔夫冈·冯·歌德(b)和罗莎琳·富兰克林(c)

牛顿所做的这个实验享誉世界。该实验的第一部分与笛卡儿(Descartes)以及其他人先前所做的相似：让太阳光穿过一个位于深色屏幕上的小孔，仅有一小束光可以穿过小孔；然后让这一小束光透过棱镜并照射在屏幕上。这时，我们就会在屏幕上看到类似彩虹的颜色带。牛顿认为这一系列颜色就是白光被分解后的颜色，且这些颜色具有普遍性。歌德被这一现象深深吸引，他从当地的一个贵族那里借来一些棱镜并自己动手做起了实验。他很快便得出结论，认为牛顿的实验完全是错误的，这是因为歌德自己发现了一套完全不同的颜色。

在歌德的实验中，他透过棱镜去看窗框。他和牛顿的做法完全相反，他在一片明亮的背景中观察一条黑线，因此，他看到的颜色和牛顿观察到的完全不同。相比于牛顿观察到的红色、绿色和蓝色，歌德观察到的则是青色、品红色和黄色，这一套颜色是牛顿所观察到的色

谱的补色。将牛顿看到的颜色合在一起得到的是白色，将歌德看到的颜色合在一起得到的则是黑色。

歌德认为，颜色是我们感知到的一种事物，牛顿却认为这是光的一种固有特性。他们的观点其实都是正确的。今天，我们已经将颜色的物理属性与它的生理学特性(对颜色的感觉)分离开来了。个体对颜色的反应是各不相同的。事实上，基于这个原理，彩光甚至可以用来进行医学治疗。从艺术的角度来看，我们的意识脱胎于对某种特定颜色光的感知，这种解读十分重要，可以简单地理解成感知颜色是非常重要的。然而，从物理学角度来看，我们可以明确地为“颜色”这一标签赋予一个基本物理特性——频率，至少在我们进入量子光学领域之前可以这么做。

光的范围远远超出了可见光谱的范围。从蓝色可见光的一端向不可见光区域扩展，会依次经过紫外线、远紫外线谱区，接着延伸至X射线、γ 射线谱区。从红色可见光的一端向反方向扩展，则会依次经过红外线、微波、无线电波直至T射线[3]谱区(见图4)。要“看到”它们，仅仅用肉眼是不够的，我们还需要借助其他各式各样的工具，但至少我们已经知道这些“颜色”的光是存在的。例如，我们之所以能够感受到太阳的温度，是由于我们的皮肤吸收了太阳辐射出来的红外线；低频微波常常被用于手机通信，还可以通过加热食物中的水分来烹饪食物。波长较短的不可见光也很常见，例如太阳辐射出的紫外线会导致皮肤晒伤，而X射线常常被用于医学成像。

X射线也被应用在很多非医疗领域，例如可以利用X射线衍射图揭示分子或者固体的结构。将X射线照射到分子或者固体中时，如果其构成原子是规则排列的，那么经过这些原子散射出来的X射线会形成一定的图案。即使原子之间的距离是人类头发丝直径的万分之一，我们也可以从该图案中推断出原子的排列结构。最著名的例子也许就是半个

多世纪前，詹姆斯·沃森(James Watson)和弗朗西斯·克里克(Francis Crick)根据罗莎琳·富兰克林(Rosalind Franklin)(见图3c)与莫里斯·威尔金斯(Maurice Wilkins)拍到的X射线衍射图确定了DNA的分子结构。这一发现让我们了解了分子的复制机制，为生物医学领域带来了巨大的变革。

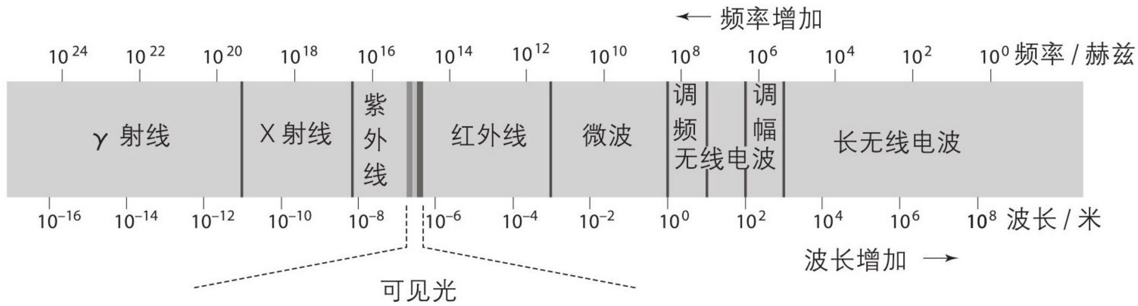


图4　电磁波谱

这些应用都显示了光的重要性。从广义上讲，光帮助我们构建起现代世界，并让我们能够充分享受现代科技。这一切都仰仗19世纪许多科学家的基础研究工作，他们是：迈克尔·法拉第(Michael Faraday)、汉斯·克里斯蒂安·奥斯特(Hans Christian Oersted)、安德烈·马利·安培(AndréMarie Ampère)、查尔斯·奥古斯丁·德·库仑(Charles Augustin de Coulomb)、亚历山德罗·伏特(Alessandro Volta)、乔治·欧姆(Georg Ohm)、詹姆斯·克拉克·麦克斯韦和海因里希·赫兹。可见光与其他与之大相径庭的不可见光，如微波和X射线之间存在着某种联系，这种联系非常引人注目。能够发现这种联系可以说是这些科学家以及其他贡献者取得的一大胜利。

色域，或者称为光谱，为艺术和科学提供了工具。画家或者艺术家致力于探索如何对各种色彩进行组合，而光谱学家则注重探索物质对不同颜色的反应。例如，在19世纪早期，约瑟夫·冯·弗劳恩霍夫

(Joseph von Fraunhofer)通过观察太阳辐射出的某些特定颜色的光，从而确定了太阳存在的一些原子类型。他留意到太阳光谱中缺失了一些特征颜色，并指出这些颜色是某些特定原子的“指纹”。光谱研究是光谱学的领域之一，它利用光来识别不同的原子和分子。现在，光谱学是一项很重要的学科，对从健康监测到远程检测大气污染物等多个领域都有着重大影响。

除了这些为人们所熟知的性质，我还想指出光的另外一个性质。它也存在于我们日常生活中的各个方面，只不过没有光的其他性质表现得那么明显。它就是光的偏振性。

3D电影就是利用了光的偏振性。看3D电影需要观众戴上特殊的眼镜，眼镜的框架是硬纸板或者塑料，框架内夹着塑料片做的“镜片”。如果你拿两副这样的眼镜，把其中一副的左镜片平移到另一副的右镜片上，并透过这两个交叠的镜片去观察一个发光的灯泡，灯泡看起来会非常昏暗。或者，你可以把其中的一副眼镜相对于另外一副旋转90°，并把两副眼镜的左镜片与左镜片(或右镜片与右镜片)重合，也可以观察到同样的现象。也就是说，光几乎不能透过如此交叠的两个镜片。

这个现象可以利用光的“方向”性进行解释。常见光源所发出的光并没有什么首选的传播方向。当你透过这种有调光作用的镜片观察一束光时，会发现光变得暗淡了一些，这是因为镜片选择了特定的光传播方向。左镜片允许某一个方向的光透过，右镜片允许的透光方向则与其相互垂直。这就是为什么当你将第二个镜片旋转90°并与第一个镜片对齐时，没有光可以透过：因为通过第一个镜片的光具有方向性，但是该方向并不是第二个镜片允许通过的方向。这种方向性特征就称为光的偏振。科学家提出并理解偏振这个概念经历了大量细致的

探索。偏振是基于光的应用的一个重要特征，对理解光的本质而言非常重要。

强度、颜色和偏振这些物理特性使得光可以用来辨别、测量和控制物质。基于这些特性，一系列工具得以发明，从而实现对物质甚至尺寸更小的对象进行研究和操控。在这一章所举的例子中，光几乎都扮演着信息载体的角色。无论是一张图片、一段光谱还是一次电话通话，光都扮演着“信差”的角色。除此之外，光在其中还有其他的一些用途。例如，我们可以利用光的热效应对金属和其他材料进行精确切割。相比于用锯子，用高能激光加工厚达几厘米的金属片会更快、质量更好且浪费更少。在医药学方面，光也有着多种用途，从激光手术矫正视力到激活抗癌药物都有涉足。

光使我们能够在任意可想象的时间和空间维度上观察自然界。在时间维度上，我们既可以观测到宇宙形成的初始时刻，也可以观察到电子在原子、分子内部以难以想象的超高速运动。在空间维度上，大至宇宙中星云的排布，小至石墨烯中碳原子的排列，都可以被观测到。光还帮助我们深入了解自然界赖以存在的基础，从量子物理学中的奇怪现象到DNA分子的结构，不胜枚举。

纵观光学的发展历程，我们可以看到人类对光的新发现使新技术得以应用，这些新技术反过来又促使许多科学领域产生新的发现。从眼镜的发明，到如今最精确的原子钟，再到现代成像、测试、通信技术，光在每一个阶段都有着不同的应用，为我们的生活方式带来了革命性的变化。尽管光学是一门古老的学科，但是这个新发现与创新所形成的循环使得光学依旧散发着蓬勃的生命力。本书将首先介绍我们如何一步一步形成光的现代理论，接着讨论我们如何使用光，产生新的理解世界的方式和新的改变世界的能力。

[1] 勒克斯(照度单位，法定符号lx)。

[2]　流明(光通量单位，法定符号lm)。

[3]　即太赫兹射线。

02 光线 Light Rays

自拍的时候，你会很自然地将手机摄像头正对着自己，这样才能保证你的影像出现在照片中。这个我们习以为常的动作表明了光的本质：物体成像时要求物体(比如说自拍时候的你)必须与相机镜头连成一条直线。这条直线通常被称为“视线”。因此，光是一种由物体向观察者直线传播的物质。

事实上，这与我们对某些类型光源的了解完全一致。音乐会上引人注目的视觉效果是由激光制造出来的，五颜六色的光束被用来照亮舞台与表演者。激光笔则通常用在课堂上，以强调屏幕上的图像或文字。这些由相干光源产生的光束高度聚焦，即使穿越整个音乐厅也几乎不发散。这表明，激光是沿直线传播的——你将激光发射器指向哪个方向，它就会往哪个方向传播。

然而，太阳光并没有明显地表现出这种特性。这也是为什么古时候人们需要费一番周折才能解释这种现象：即使两个物体的物理尺寸完全相同，远处的那个物体看起来也要比近处的物体小。如果知道了光沿直线传播的特性，这个现象就非常好理解了。

在公元前300年左右，古希腊的欧几里得就已经用光的这一特性来解释“近大远小”的现象了，他的想法被记录进了最早的光学书籍之一[1]。如图5所示，想象两条直线，用来表示从物体传播到观察者的光线，一条将物体(图5中的支柱)的顶部与观察者的眼睛连接起来，另一条则是从物体的底部连接到眼睛。这两条线之间的夹角越大，我们感官上就会觉得像越大，反之同理。尽管离观察者较远的支柱与较近的支柱物理尺寸完全相同，但由于它与观察者间的连线夹角相比来说更小，因而远处的支柱看起来更小。我们把图5这样的图称为透视图。

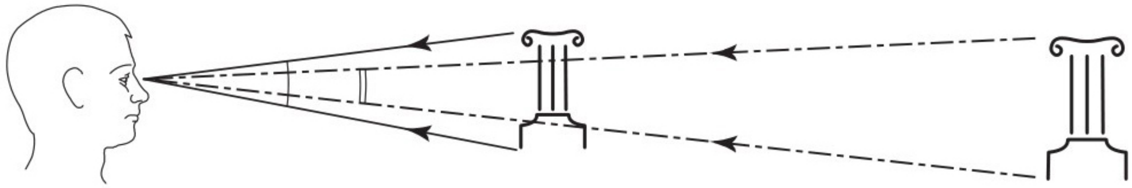


图5　欧几里得用光线沿直线传播的原理解释了为什么相同尺寸的物体在离观察者更远时看起来更小

那么，构成光线的物质是什么？基于更早期的学说，欧几里得认为这种物质是从眼睛里发出的粒子，它来自想象中的体内火焰，照亮物体并被物体反射回观察者的眼中。如果这个说法成立，就表明不管外界是光明还是黑暗我们都能够看到东西，这显然与我们的生活经验相违背。尽管如此，粒子在物体和观察者之间沿着某一轨迹移动的想法仍然是一个强有力的学说。

11世纪的阿拉伯科学家海什木修正了欧几里得的理论，从而逐渐形成我们现在所确信的理论：物体被来自太阳的光线(想象中的外部火焰，与体内火焰相对应)照亮，光线再通过反射传送到观察者。至于海什木是如何想到这个理论的，流传着好几个故事，其中一个是关于他做的直视太阳的实验。直视太阳时眼睛会灼痛，他认为，如果构成光线的物质是由眼睛发出的，即“体内火焰”一直在燃烧，那么不管他有没有直视太阳，这种灼痛感会一直存在。因此他认为，物体成像所需的光源来自外部，而不是由眼睛发出的。

在这个理论基础上，我们假设沿光线移动的是光的粒子，称为**光子**。于是，光束的亮度就与一秒内光线中通过的光子数有关。为了理解物体是如何成像的，我们需要考虑当一个光子在镜面上反射以及穿过透镜时将会发生的情况，从而推导出“光学定律”。运用光学定律，人们可以设计出非常复杂的光学仪器，例如：手术显微镜，用于

“锁孔”手术[2]的导管，以及放置在绕地轨道上用于观测遥远星系的大型光学望远镜。这些仪器深刻地影响着我们的生活和我们对世界的理解。

这些光粒子具有什么样的特性呢？粒子的常见属性包括其位置、行进方向和速度信息。我们作出如下假设并暂不深究：光粒子以光速移动；光子的位置可以指定为光线的起始位置；光子的运动方向则是光线的方向。在这些假设下，光子可以看作以光速从起始点射出，沿着光线的方向运动直到它遇到了某一物体的表面。

反射

当光传播到物体表面时会发生反射现象：光子从物体表面反弹，进而改变其运动方向，但这并不改变它在物体表面的位置，如图6所示。早在公元1世纪，亚历山大城的海伦(Hero)就已经发现了光的“反射定律”，阐明了光子运动方向改变的规律：入射角(入射光线方向与入射点垂直于表面的方向之间的夹角)等于反射角(反射光线方向和入射点垂直于表面的方向之间的夹角)。这个定律在概念上十分简单却非常有力，我们可以用它来解释非常多的光学现象。

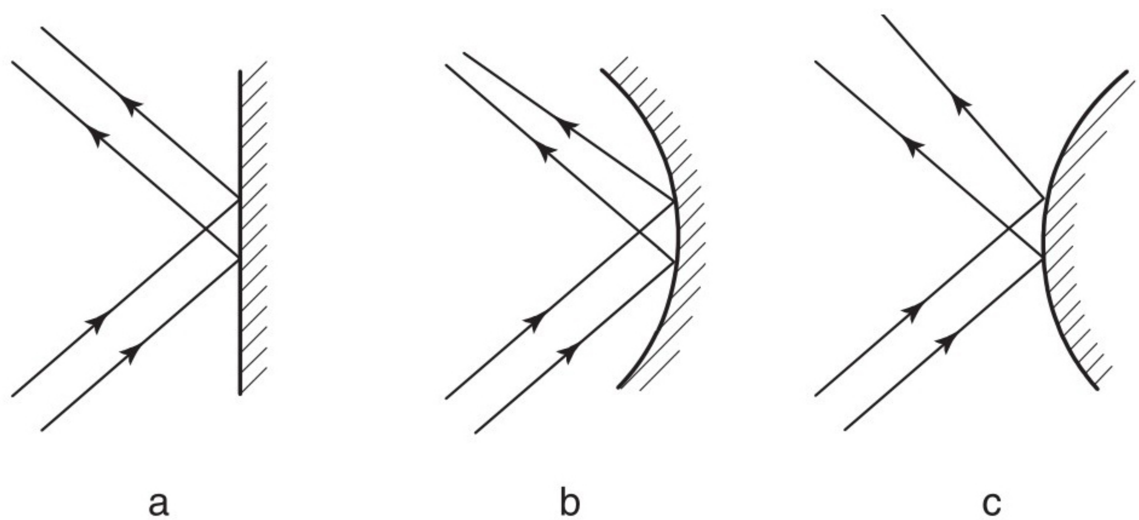


图6　光线分别反射在平面镜面(a)和曲面镜面(b和c)上

让我们先来看看由反射造成的左右反转现象。如果你把手表放到镜子前并观察它的镜像，你会看到镜中的秒针正沿着逆时针方向移动；当你移动你的右手时，镜中的你却在移动左手。这种“手性”[3]的变化是镜像的标志——镜中反映的世界在这种意义上是真正的左右颠倒。

这一现象完全可以通过海伦的反射定律来解释。图7显示了镜面如何使镜像发生了“手性”的颠倒。图中，顺时针方向运动的箭头是我们的观察对象，箭头上每个点出发的光线同时由平面镜反射并重新排列，使得镜子中箭头的镜像指向逆时针方向。你也可以使用相同的方法来解释为什么指向左方的箭头其镜像指向右方，反之亦然。但是，向上箭头的镜像仍然指向上方，向下箭头的镜像仍然指向下方。

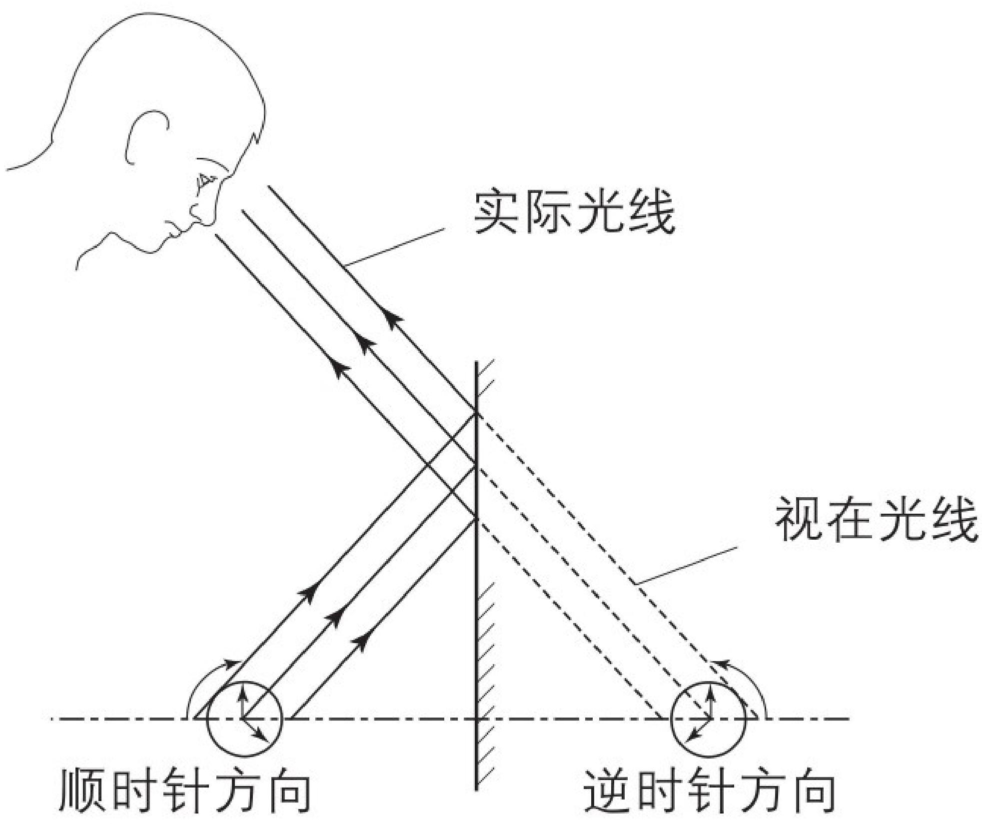


图7　顺时针旋转的时钟指针在镜子中变为逆时针旋转的光路示意图

反射成像

在浴室的平面镜子里，你可以看到左右反转的自己，但如果你对着一把抛光的勺子观察自己的反射图像，你会看到自己的形象被扭曲了：你的面部特征在勺子凹陷的弧形表面上被放大了。事实上，勺子凹陷的前表面有放大物体的效果，凸起的后表面则有缩小物体的效果。

为什么会出现这样的现象呢？从用于视力矫正的隐形眼镜到用于科学发现的太空望远镜，这些光学仪器的制造都运用了光的成像原理，因此弄清楚光是如何成像的非常重要。

到目前为止，我们只考虑了物体上某一点的某一道光线。实际上，光线通常是从物体上的各个点向四面八方散射的。如图8所示，假设多条光线组成的“光线束”从物体上的某个点出发并形成一个锥体，这个光锥在远离物体时发散。这些光线将在弧形镜面上的不同点处发生反射，因此形成了不同大小的入射角。尽管光线的反射方向不相同，但是在每个方向上的反射现象仍然满足海伦的反射定律。反射之后的光线会形成一个逆向的光锥，并最终会聚在同一个点上，这个点就是对应发出光线束的原始物体点的“像点”。

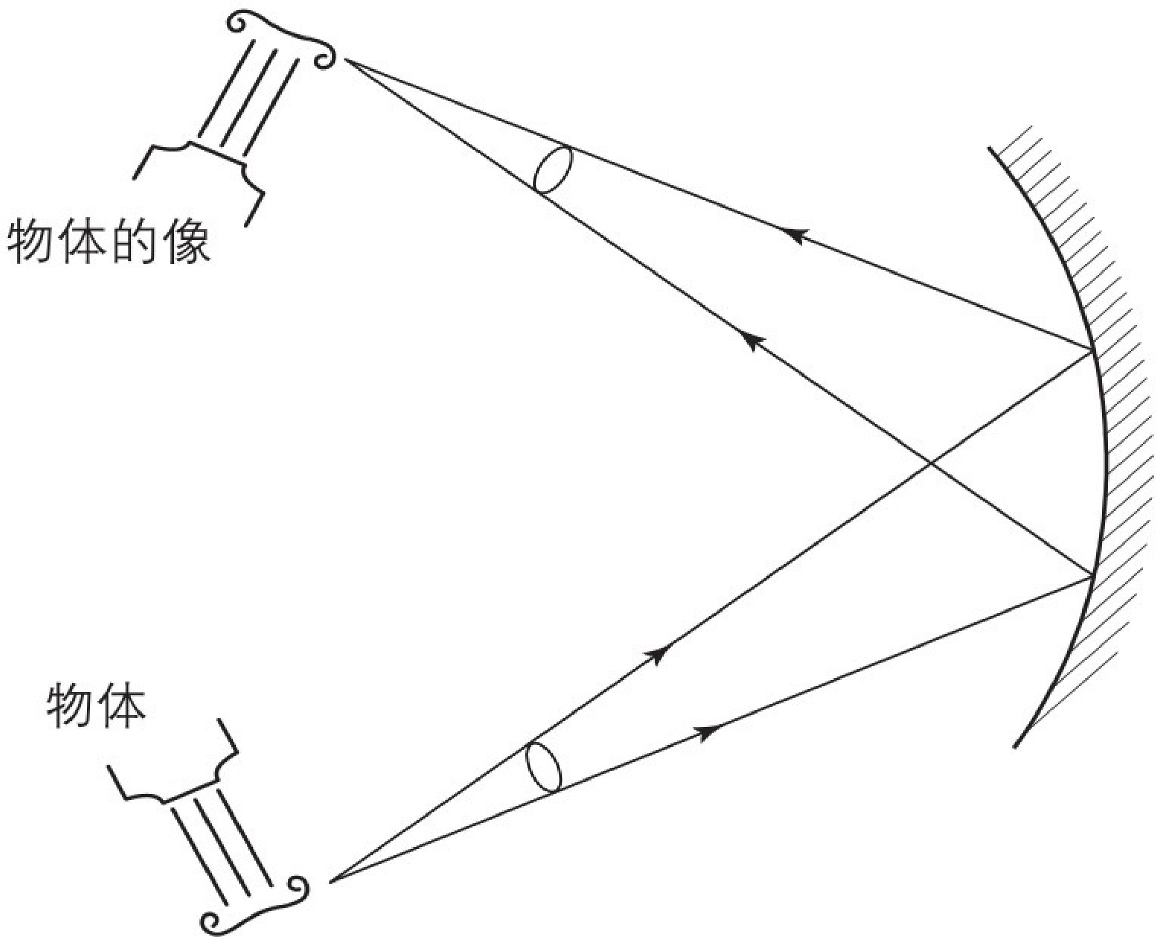


图8　物体通过曲面镜成像示意图。物体上某一点发出的锥形光束通过曲面镜的反射最终会聚到一个像点上

我们通常认为，物体的像是由物体上各个点所对应的像点组成的。像的大小由物体与曲面镜的距离以及曲面镜的聚焦能力决定，其中后者由曲面镜表面的曲率半径决定(例如，凹陷弧度越大的曲面镜表面曲率半径越小)。当物体更靠近镜子时，像可以比物体本身更大。像与原始物体之间大小的比率被称为放大率。

牛顿利用曲面镜可以放大成像的特性来设计望远镜，如图9所示。他的设计有一个显著的特点：不论是什么颜色的光线，望远镜对远处

物体成像的大小是不变的(没有色差)。显然，牛顿巧妙地利用了反射定律里入射角必然与反射角相等的特性，使得无论光线的颜色是什么，只要光线的入射角度相同，反射角度就一定会相同。因此，每种颜色的像都将形成于同一位置，这就使所有颜色得以完美保留。



图9　牛顿的反射望远镜，它的成像没有色差

折射

牛顿之所以发明了这种反射望远镜，是因为伽利略·伽利雷和约翰尼斯·开普勒(Johannes Kepler)等当代先驱所使用的望远镜受到了色差的严重影响。他们的望远镜成像时，在观测物体边缘总是有一个模糊的彩色光环，其原因在于这些望远镜是利用光的折射特性设计的。光的折射指的是当光从一种透明介质传递到另一种透明介质时传播方向发生弯曲的现象。

正是由于光的折射现象，使得浸入水中的铅笔看起来好像沿着水面被“折断”了。这就是光的折射定律，通常被称为斯涅尔定律，以荷兰物理学家威理博·斯涅尔(Willebrord Snell)的名字命名，他在17世纪早期就发现了这一定律。该定律表明，折射光线与入射点垂直于表面的方向间的夹角不仅与入射角有关，还与两个透明介质的特性有关。如图10所示，在铅笔看似被折断的例子中，这两个介质分别是空气与水。

“折射率”被用来度量透明介质的特性，它的大小反映了相应介质光学“刚度”的强弱。例如，光在具有较大折射率的介质中传播得更慢，这是因为光线更不容易“挪动”该介质分子中的原子与电子，即具有较大折射率的介质对光的阻力更大一些，我们可以理解成它的光学刚度更高一些。这就像在水池中奔跑，如果水不深，你的腿可以轻松移动；但如果水深达膝盖，你就没那么容易在水下自由行走了，因为你必须抵抗水的阻力。

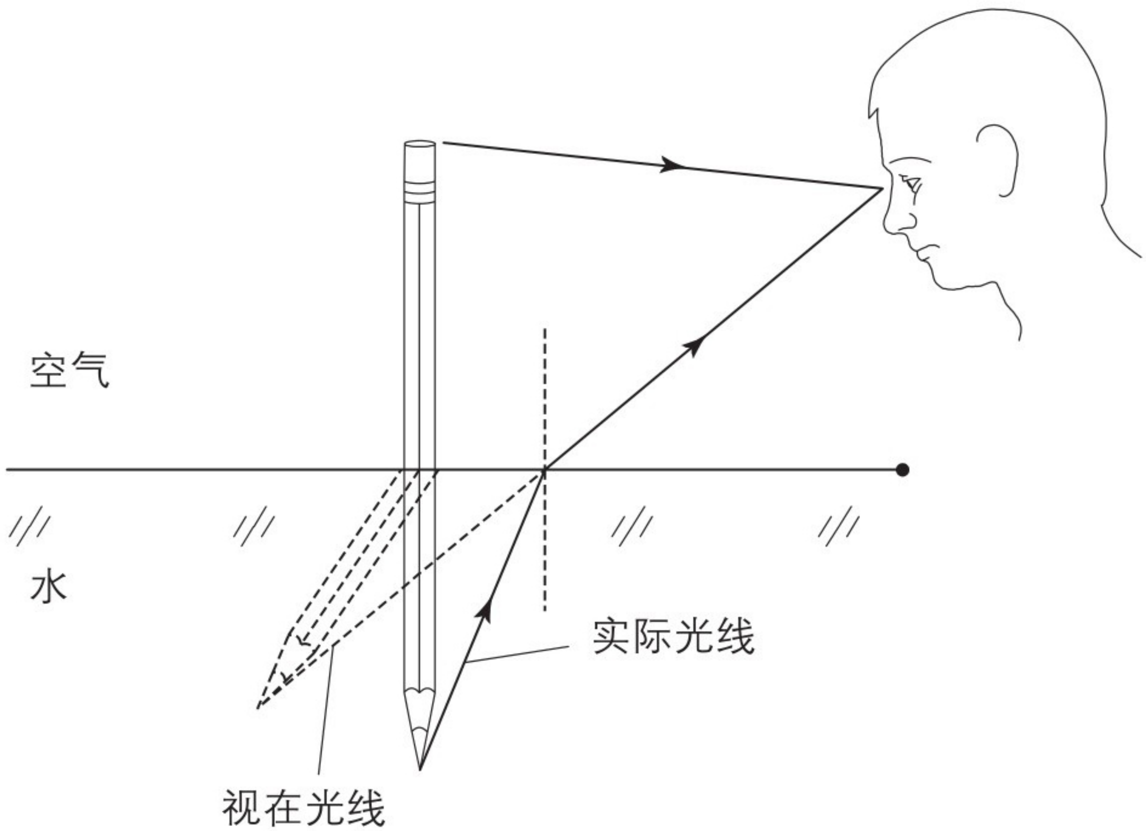


图10　光线在空气和水的交界面上发生折射

事实上，折射定律还可以用另一种方法推导。皮埃尔・德・费马(Pierre de Fermat)表明，当光从一种介质中的某个点传播到另一种介质中的某个点期间，它会寻找到一条特别的传播路径，使得光能用较短时间从高折射率的介质中穿过，用较长时间从低折射率的介质中穿过。这就要求光线在两种介质的交界面处发生弯曲，即著名的费马原理，它与斯涅尔定律异曲同工。

透镜成像

我们已经知道，光可以通过有弧度的镜面反射成像，与此类似，光也可以通过有弧度的透镜折射成像，其成像过程如图11所示，来自物体上某点的一束光线通过透镜的折射最终聚焦在像的一个像点上。

请注意图中透镜的形状，它的横截面是不是很像扁豆(lentil)？这就是透镜(lens)这个词的来源。

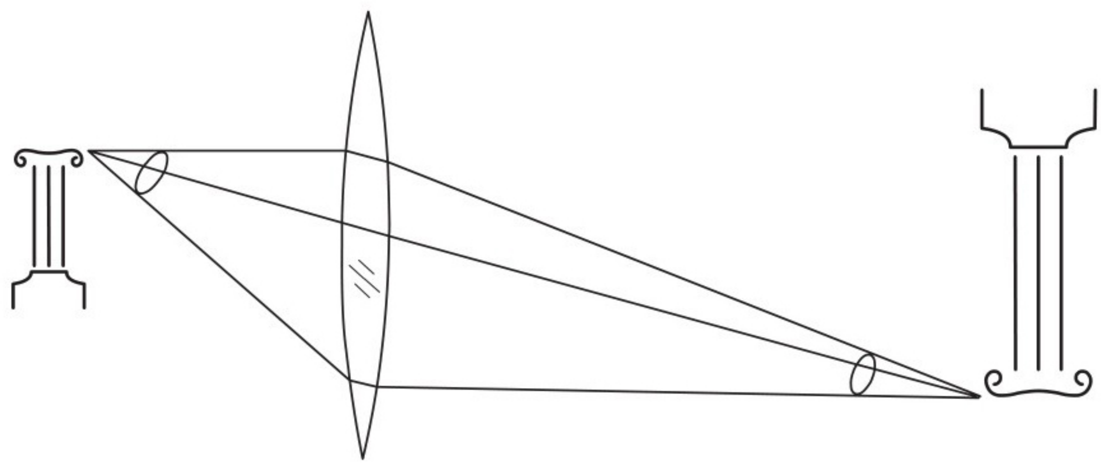


图11　光线通过透镜成像示意图

从你的眼睛到手机摄像头，再到手术显微镜，透镜在成像设备中无处不在。成像仪器具有两个部件：透镜本身和光学检测仪。光学检测仪是将光转换成其他信号的仪器，通常是电信号。例如，视网膜就是眼睛的光学检测仪，而手机摄像头的光学检测仪则是固态光传感器，由一系列微小的硅片组成。

诚然，每个装置中的透镜都不尽相同，但基本原理是一样的。不论在哪种情况下，镜头与光学检测仪之间的距离都是关键的设计参数；另一个关键参数则是透镜的**焦距**，它可以衡量透镜“弯曲光线”的能力。焦距由透镜的曲率以及厚度决定。要制造短焦距的透镜，需要采用曲度大且更厚的材料。这样的透镜通常用于需要高放大率的仪器，例如显微镜。

光的颜色改变，透镜材料的折射率也会随之改变，所以不同颜色的光在透镜表面会有不同程度的弯曲，导致每种颜色的焦点出现在不

同的位置。这使得透镜成像的周围会出现不同颜色的“晕圈”。例如，在一个特定的光学检测仪平面，通常只有一种颜色可以准确聚焦，而其他颜色将失焦并形成光环。这种色差现象是否会导致严重后果需要具体情况具体分析。

我们最熟悉且最重要的成像工具之一就是我们的眼睛。它由前折射表面、角膜和可调节透镜依次组成，其中可调节透镜可以根据眼睛聚焦物体的远近而改变形状。物体通过这一系列眼部结构后，最终在眼睛后部的视网膜上形成图像。

在历史上，眼睛成像的原理一直令人们非常感兴趣，尤其是在笛卡儿做了一个关于眼睛成像的实验之后(见图12)。实验显示直立物体的图像经过眼睛成像后会上下颠倒。当然，我们眼睛看到的并不是上下颠倒的物体，很明显，大脑一定对原始视网膜信号进行了一些非凡的处理，校正了其与外部世界不一致的部分，使我们的感知变得准确。

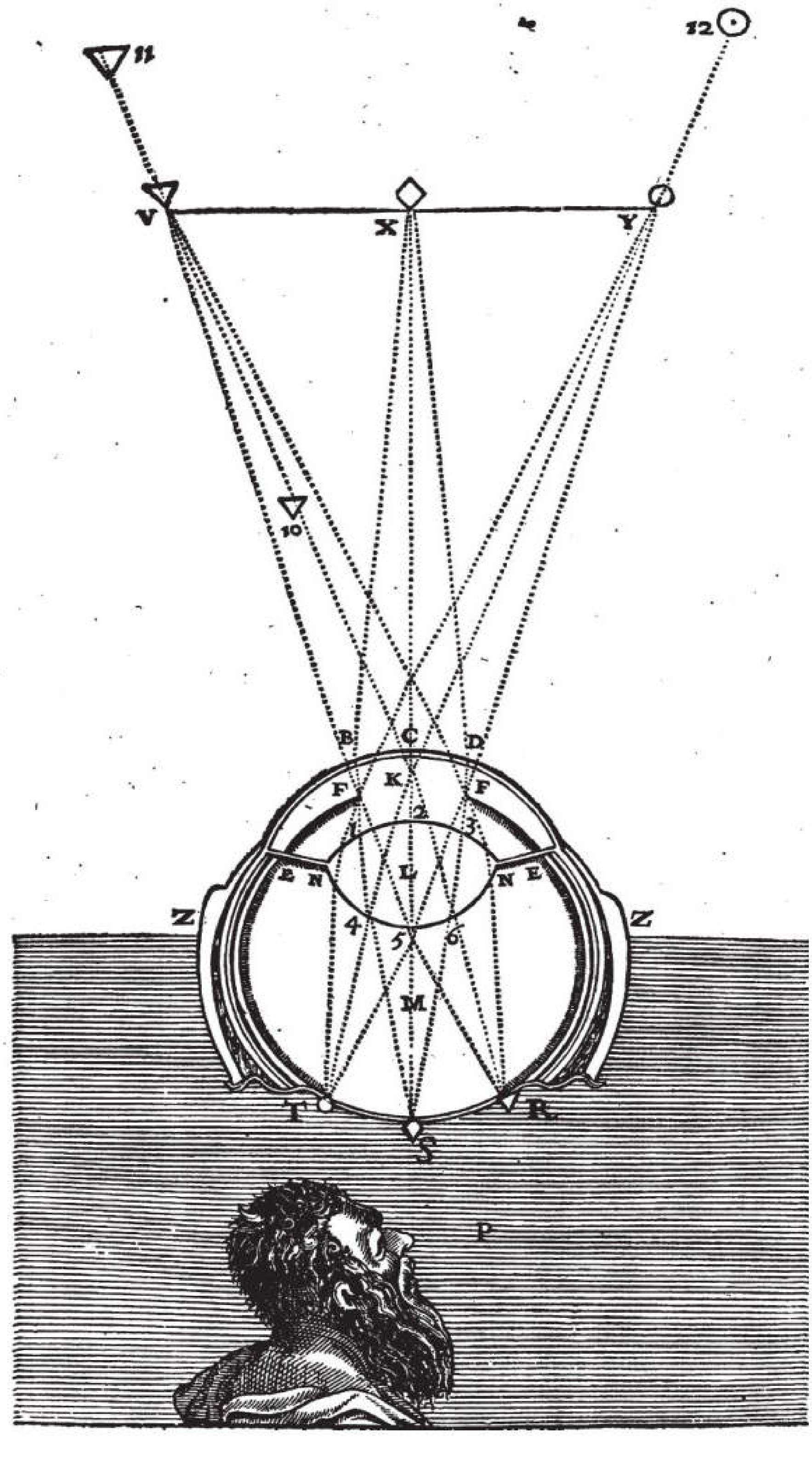

11
12
V
X
Y
10
B
C
D
F
K
F
1
2
3
E N
L
N E
Z
Z
4
5
6
M
T
S
R
P

图12　笛卡儿关于眼睛成像的实验。实验显示眼底视网膜成的像是上下颠倒的[4]

光学仪器

众所周知，眼睛看清东西的能力(形状清晰、色彩鲜明)会随着年龄的增长而下降。最早的一些光学仪器就是开发出来作为视觉辅助的。眼镜可能是第一个这样的光学仪器，据称是由罗杰·培根(Roger Bacon)在13世纪时发明的，他被称作牛津的“疯狂修道士(mad friar)”。

眼镜通常是将简单的透镜，镶以镜框，佩戴在距离角膜(眼球的前表面)一定距离(通常是几毫米)的位置。“隐形眼镜(contact lenses)”，顾名思义，则是将透镜直接与角膜接触(contact)。在这两种情况下，成像系统都是复合的，也就是说，成像系统由外部透镜、角膜和晶状体这几个部分组成。这样就可以通过外部透镜去补偿眼睛晶状体的缺陷，从而达到矫正视力的目的。这种矫正也可以通过激光手术直接改变眼睛前表面的形状来完成。激光辅助原位角膜磨削术(Laser-assisted Subepithelial Keratomileusis，缩写为LASIK)就是这样一种手术，它使用激光烧蚀角膜表面的一部分以改变其曲率，从而改变角膜的聚焦能力，即改变了眼睛的成像能力。

许多其他成像工具的工作原理与眼睛非常相似，例如手机摄像头。手机摄像头通常非常小，但却能够拍出高质量的图像，可供我们在社交媒体上发布。手机的摄像头被放置在手机表面，硅基光电探测器阵列则被放置在手机内部。手机摄像能否得到高质量图像取决于两个方面：一是阵列中探测器的大小和数量，二是光学系统创建无色差、清晰、无失真图像的能力。

探测器阵列的“像素”数量大小一般用来描述图像质量的好坏：一个2400万像素的摄像头(探测器阵列包含2400万个传感器)通常被认为比800万像素的摄像头更好。像素可以被认为是最小可成像单位的图像大小。当探测器阵列中传感器数量很少时，被拍摄物体就只能被解析成少量的最小可成像单位，也就是说像素量很少，那么从拍出来的图像里就很难分辨出这个物体。因此，像素一般是越大越好，但前提是成像系统能够产生的最小图像单位比探测器元件还要小。

成像极限

在19世纪，德国科学家恩斯特·阿贝(Ernst Abbe)设计了一个简单的规则来描述成像极限，即像素尺寸大小。该规则适用于任何当时已知的成像系统。在阿贝准则中，一个像素的大小()与照射该物体的光的波长(λ)乘以透镜的焦距()再除以镜头直径()的结果成正比。

$$S = 1.22 \times \lambda f / D$$

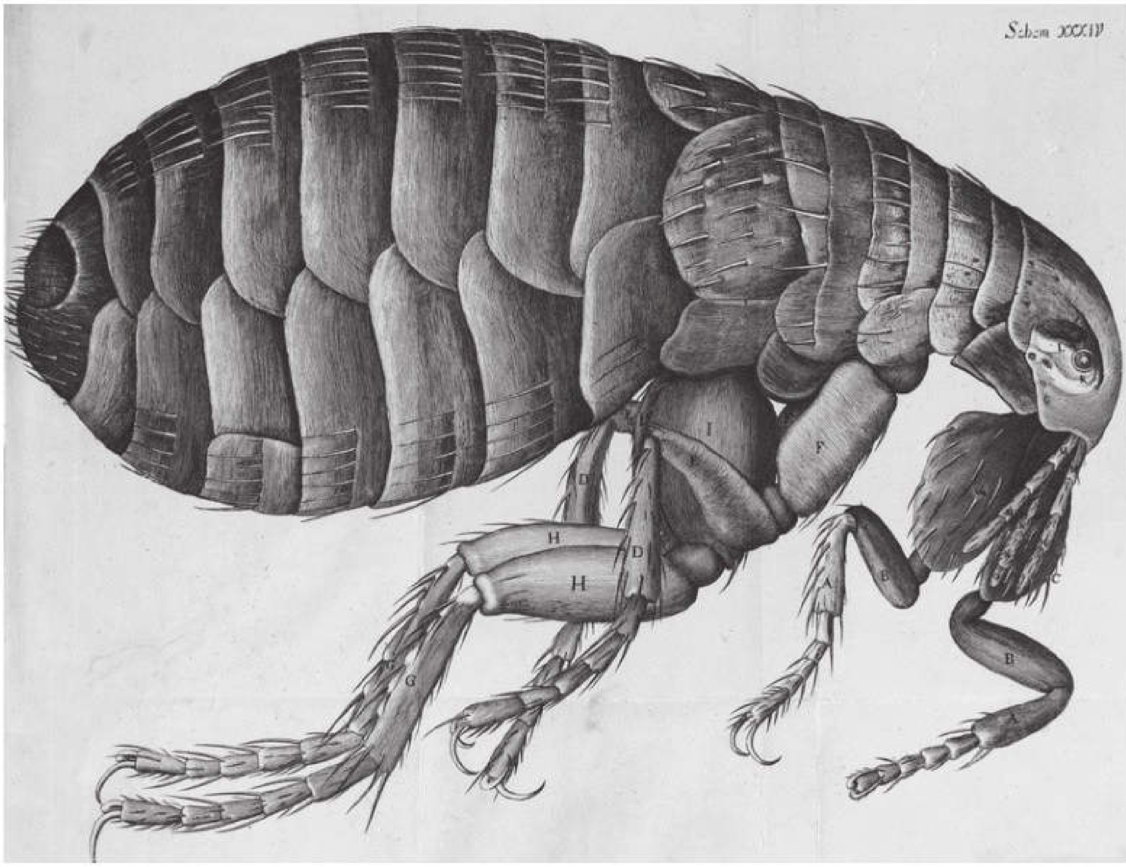
因此，具有大直径和短焦距的透镜将会有更小的像素尺寸，因而成像更为清晰。在所有透镜系统中，像素大小与物体的光的波长相同时得到的物体图像，就是我们能得到的最佳图像。也就是说，实现最佳图像效果时像素大小大约为一个波长，这也是大多数光学仪器(例如相机和双筒望远镜)像素尺寸的极限。

在光学的许多重要应用中，设计和构建能够产出高质量图像的成像系统一直是重中之重。例如显微镜，它被广泛应用于从生物学研究到外科手术的多个领域中。最早的显微镜使用的是非常简单的透镜，尽管它很小，具有类似于球形抛光玻璃的形状，却为17世纪的罗伯特·胡克(Robert Hooke)等早期实验者提供了探索自然界中无法用肉眼

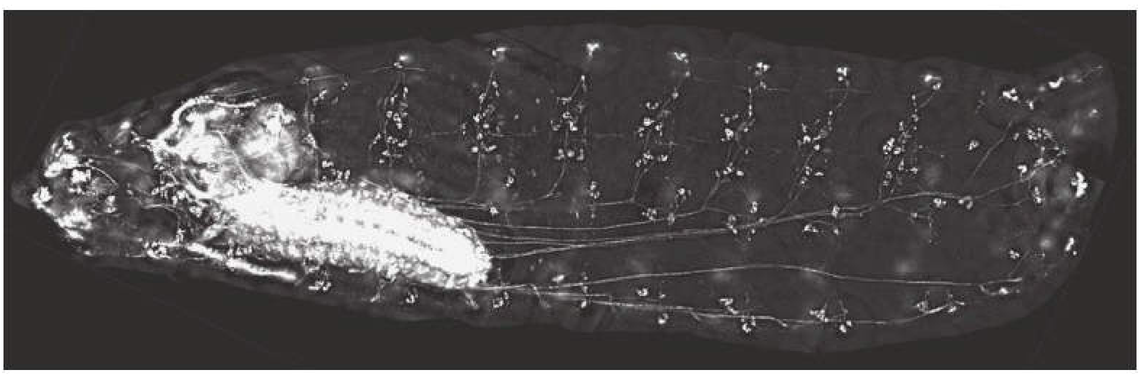
看到的微小生物的工具。如图13a所示，胡克绘制的跳蚤图揭示了显微技术的力量，使更多的新发现成为可能。

相较起来，现代的科研显微镜设备更为复杂。它们由包含多个部件的复合透镜组成，使得成像的像素大小非常接近光的波长，即之前提到的成像极限，或阿贝极限。图13b显示了一个由现代显微镜成像的例子。这是一张果蝇幼虫神经系统的合成图像，该幼虫即将孵化。它是通过位于幼虫细胞中的发光蛋白成像而成的。

阿贝极限适用于图像亮度与物体亮度成比例的光学系统，这样的光学系统被称为线性系统。事实上，阿贝极限可以被突破，但是需要通过非线性系统，在这种系统中，图像亮度与物体亮度的平方乃至更复杂的函数成比例。为了更全面地解释这些效应，需要更多地了解光的波动模型，这将是本书第3章的主题。



a



b

图13　a.胡克通过早期显微镜观察到的跳蚤图；b.用现代荧光显微镜拍摄的果蝇幼虫的神经系统

具有类似特性及复杂性的光学成像系统还被用于计算机芯片的制造中。要知道，电子电路元件都非常微小，连接芯片上两个晶体管的导线直径仅为250纳米。复杂的器件和连接阵列通过一种被称为光刻的工艺布置在硅片上。

为使设计师观察方便，芯片布局的绘制比例一般都足够大，绘制完成后，将芯片布局按比例缩小并投影到芯片上；接着，这一图像被蚀刻到晶片的表面涂层上；最后，通过一系列化学反应，图像被映射到实际设备上。要完成这一系列步骤，成像系统必须具有非常高的分辨率，即像素尺寸要与设备连接线的尺寸数量级相同。在整个晶片上保持这种分辨率是一个很大的挑战，这需要对许多透镜元件精密组合，从而将所有像差减小到绝对最小值。图14就展示了这样一个例子，图中呈现了该透镜的横截面，显示了透镜元件与光线路径的多样性。

刚刚我们讨论了显微镜，现在我们来谈谈另一个极端——天文望远镜，它们有的在地面，有的在太空，尽管都是非常大的仪器，却是由相对简单的光学成像元件构成，通常只用一个弧形的反射表面和一个简单的“目镜”来调整光线，以便与现有探测器充分匹配。这些成像系统的显著特征就是它们的尺寸。

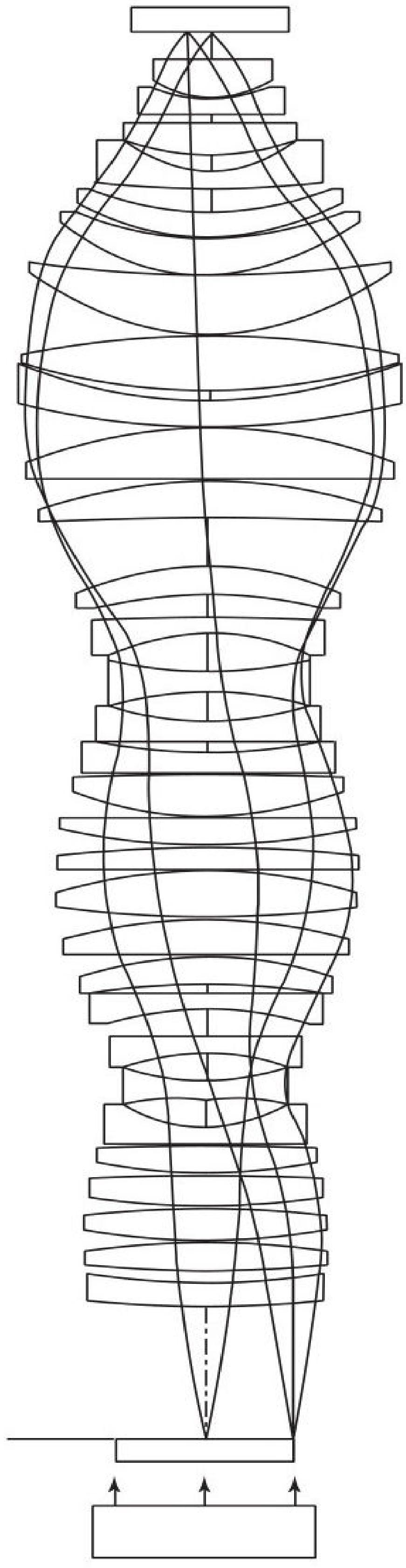


图14　用于计算机芯片光刻的透镜的一部分。它由20多个不同的透镜元件组成，可以形成500纳米的图像，这个尺寸比其使用的光的波长的一半还要小

当我们观测非常遥远的恒星时，它们看起来十分微弱，这是因为它们发出的光线几乎都没能到达地球。因此，尽可能多地收集这些光线是非常重要的，这往往需要一个非常大的透镜或镜面——直径达几十米或更大。制造这种尺寸的透镜是不切实际的，但镜面可以。因此，巨大的镜面被制造出来并安装在大型望远镜中。同时，为了收集足够多的光线来成像，往往需要长时间观察遥远的恒星。这导致了地面望远镜的另一个问题：大气层不是一成不变的，它的密度随着风、温度和湿度变化，这些波动往往会使光线偏离其传播方向。比如，恒星之所以看起来在“闪烁”，就是因为大气湍流使光线随机照向或偏离望远镜的探测器。

解决这个问题的方法之一就是直接将望远镜放在大气层外的太空中，哈勃太空望远镜就是一个例子。它成功获取了遥远恒星、星系和星云的壮观图像，观测到遥远太空中非凡的宇宙结构与运动。除此之外，还有其他的解决方案吗？光学工程师在过去20年中设计了一种巧妙的方法，为地面上的可见光望远镜成功解决了这个问题：将望远镜的镜面分割成多个区域，每个区域的镜面都可以倾斜，通过调节镜面不同区域的倾斜角度，就可以“操纵”光线，使它们全部被探测器所接收。如果你可以测量光线穿过大气层时产生的偏差，就可以通过调整镜面来补偿这一偏差。光学工程师们首先测量导星(位于上层大气中的人造光源)的光通过大气层的扭曲程度，然后根据这一信息来调整不同区域镜面的倾斜角度。用这种方式，地面望远镜成的像可以刚好达到阿贝极限。但是，将望远镜放置在太空仍然是必要的，这是因为有一些电磁波段会被大气吸收，例如X射线和紫外线，为了对它们进行观测，我们仍然需要太空望远镜。美国国家航空航天局(The National

Aeronautics and Space Administration，NASA)和欧洲航天局(European Space Agency，缩写为ESA)正在计划进行新的太空望远镜任务。

超颖材料和超级透镜

多年来，光学科学家们一直致力于建立卓越的光学系统，那么，是否存在这样一个具有完美的成像能力的透镜呢？从19世纪英国的詹姆斯·克拉克·麦克斯韦到20世纪苏联的维克托·韦谢拉戈(Victor Veselago)，许多伟大的物理学家都对这一问题充满了兴趣。韦谢拉戈设想了这样一种奇特的材料：当光线射到这种材料表面后并不会遵循斯涅尔定律，甚至与其完全相悖。斯涅尔定律是基于常见的“普通”材料，它们的折射率为正数，而韦谢拉戈提出的材料具有“负”的折射率。这种材料由许多微小结构组成，且每一个微小结构的尺寸都小于观测光的波长。这种特殊的结构赋予了“超颖材料”不同寻常的光学特性。举一个有代表性的例子，与光线在两种普通材料之间的界面上相比，当光线在普通材料和超颖材料之间的界面上产生折射时，折射方向将完全相反。

利用超颖材料独特的折射率，我们可以通过工程设计使它将从各个方向射来的光线弯曲。这样，本来会在材料表面发生散射的入射光线将围绕着超颖材料的表面发散出去，从而使超颖材料“隐形”。事实上，英国物理学家约翰·彭德里(John Pendry)爵士表示，使用超颖材料制作隐形斗篷是完全有可能的。

超颖材料还有另一个不寻常的特性，就是能够对非常接近超颖材料的物体进行完美成像。如果使用超颖材料制造透镜，其表面可以做得非常平整，不需要像玻璃透镜那样有很大的弧度。这使得利用超颖材料制作的透镜很适合观察非常微小的物体，尤其是仅有数十纳米量

级尺寸的纳米结构体。超颖透镜可以类比为21世纪的胡克显微镜技术，也许它会开启一个新科学发现频繁涌现的新时代。

本章描述的所有成像系统都对物体进行了二维渲染[5]，毕竟我们通常对图像的理解就是平面的图片。那么，我们是不是能设想出一个可以制作三维图像的系统呢？这就需要我们对光本身有更深入的了解，我们将在第3章中就这一点展开讨论。

[1] 欧几里得的《光学》(Optics)被认为是最早的光学著作。不过中国的《墨子》一书对光学也有一些研究，如小孔成像等，比《光学》更早一些。

[2] 一种微创手术，切口比锁孔还小，使用包括光纤在内的特殊仪器和技术。

[3] 手性指一个物体不能与其镜像相重合。

[4] 图中La Dioptrique为法语，译作“屈光”，特指在眼部所发生的光的折射。

[5] 对三维物体进行二维渲染意即利用平面图像显示出三维物体的立体感。

03 光波 Waves

在第2章里，光被认为是由基本粒子组成的，并沿着明确的轨迹运动，我称之为“台球”模型。在这一模型中，光束可看作是由一个个分布紧凑的能量束集合而成的，这与光是波的观点形成鲜明的对比。事实上，光的波动说与光的粒子说的发展可谓并驾齐驱，只不过与光的粒子说相比，光的波动说经过了很多年的争论和实验之后，才被人们完全接受。

无法解释的现象？

在阳光下观察水面上漂浮的一小层薄油，你会在油层的边缘发现彩色的轮廓。正是这一观察激发了一个跨时代的想法——光是作为波运动的。牛顿是最早描述这种现象的人之一，但是这一现象对他的光粒子模型提出了挑战。为解释这一现象，牛顿势必要对他的模型进行严重的扭曲。而在海峡对岸，牛顿在光学上的竞争对手、法国科学院院长、荷兰人克里斯蒂安·惠更斯(Christian Huygens)则使用了光的波动模型来解释这一现象。这一解释被证明更加合理。因此，早在光学研究初期，波和粒子的概念就已经同时出现了。

不仅是这一种现象，还有其他的现象也不能够用光的粒子说来解释。例如17世纪中期弗朗西斯科·格里马尔迪(Francesco Grimaldi)的一些发现。他发现光线通过小孔(比如屏幕上的一个小孔)时会偏离直线。他注意到光线发散开来，且在光束的边缘看到了彩色的条纹，这种现象对于头发或者薄纱这种小物体尤为明显。他总结说，光照射在一个小或窄的物体上所形成的条纹说明，当光经过这些物体边缘的时候偏离了其原始路径。如果光真的是由沿着直线运动的粒子组成的，那么这样的固体物质肯定会投射出阴影，而不会导致光粒子偏离成奇怪的模式。

此外，令牛顿和他同时代的人都感到困惑的一个问题是，当光通过某些物体，尤其是一些晶体，例如方解石(一种自然矿物)时，会发生古怪的折射，这个现象用光的粒子说根本解释不清楚。图15中的例子就很好地体现了这个现象。用灯泡照亮一张纸上的单词“LIGHT”，且用两块方解石分别盖住单词的左右各一半。在图15a的左半边图像中，单词由一个错位成了两个，右半边图像中的单词也变成了两个，且错位的方向相反。将图中15a左半边上方的图像和右半边下方的图像结合起来，才是符合人们期望的、由纸反射的光经过晶体折射后看到的单词。而图15b似乎是由不同的折射率产生的。如图15b、15c图所示，通过在晶体上放置偏振器，可以分离由两个不同方向偏振光形成的图像。每个偏振光都有不同的折射率，这就是**双折射现象**。

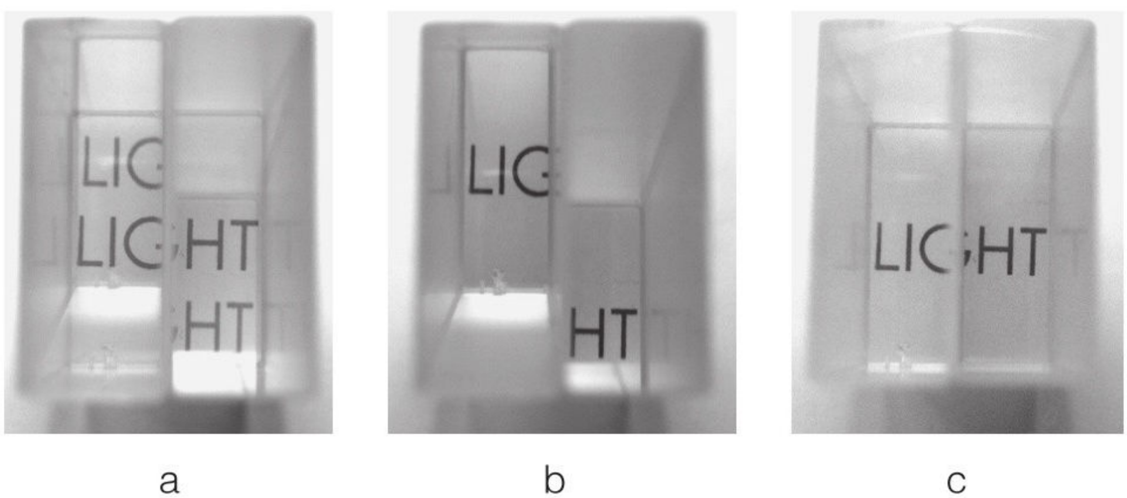


图15　展示双折射现象的一个实验。在入射光分别为非偏振光(a)、垂直偏振光(b)、水平偏振光时(c)，透过一对方解石晶体观察写在纸上的单词LIGHT

所有的这些观测结果都表明光还有一些性质无法解释。这些性质分别是**干涉**、**衍射**与**偏振**。我们将在本章对这些现象进行探索，继续讲述光的波动说。

波长和频率

波的特点是什么？波是与介质有关的一种波动形式，例如池塘表面的水波，这些波是由水分子在水与空气界面处的上下运动形成的。这种运动的最高点和最低点即为水波的波峰和波谷，而水波本身沿着池水表面运动，也就是说水波的运动方向与水分子的运动方向垂直。因此，它被称为**横波**。它的波速取决于水的深度等因素。

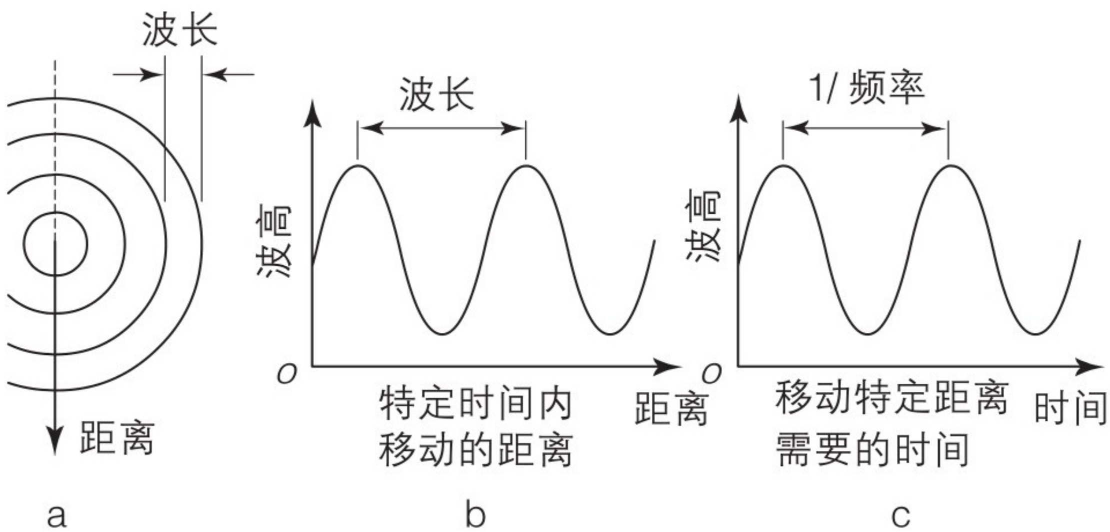


图16　水面上的圆形波：a. 等高线，也称为波前；b. 在某特定时刻，波的高度与距中心位置距离间的关系图；c. 在水面某特定位置，波的高度和时间之间的关系图

如图16a所示，一块石头被抛入水中，圆形的表面波从石头入水处产生，并往外扩散，这是一种很常见的现象。相邻波峰之间的距离称为**波长**(见图16b)，波峰到达岸边的速率称为波的**频率**(见图16c)，波长和频率的乘积称为**波速**。

很多个世纪以来，人们一直有个疑问，光究竟是由哪种波组成的？一些人认为波的存在需要某种介质。由于光的速度很大，所以这个介质必须非常坚硬，但是这样又会导致其他物体很难穿过它。比方

说，我们之所以能够看到遥远的恒星，肯定是存在某种介质使得光得以在恒星和地球之间传播。地球在围绕太阳运转的过程中必然会不断地穿入并穿出这种介质。这种神秘的介质被称为“以太”，直到19世纪末，它才被视为一个无用的概念而被抛弃。

所以，光是什么样的波呢？这个问题最终由詹姆斯·克拉克·麦克斯韦在19世纪给出了答案。他指出这是一种新的实体的振动：电磁场。电磁场是作用在电荷和磁性材料上的力。例如，一块带有静电的布会吸附灰尘颗粒，一块磁铁会被吸附到冰箱门上。在第二个例子中，当你手拿着磁铁靠近冰箱门时，你的手就能感受到这种力：磁铁会加速靠近冰箱门，除非你给磁铁施加一个反向的作用力。

在上面两个例子中，都存在一种力将一个物体拉向另一个物体。在第一个例子中，布上的电荷产生电场，因而在离布一定距离之内的灰尘将受到源自该电场的力。同样，冰箱门受到的作用力源自磁铁产生的磁场。19世纪早期，迈克尔·法拉第就已经证明了电场和磁场之间存在紧密的联系。麦克斯韦把电场和磁场合在一起，称为电磁场。在波动模型中，光可以看作电磁场的高频振荡。基于这个观点，移动的电荷可以产生光波。我将在第5章中讨论这种方法以及其他产生光的方法。

干涉

如果将两块石头扔到水里相距较近的位置，那么从这两个位置就会产生两组圆形波，它们向外扩散并最终相遇。在波与波叠加的地方，波峰会变得更高。在水面上也有另外一些可以连接成线的地方，尽管两个波都经过，却没有丝毫高低起伏。这些线的位置如图17所示，一般都是沿着波前(波源发出的振动经相同时间所到达的各个波峰点组成的面，见图17a中灰色的同心圆线，圆心则为波源位置)。

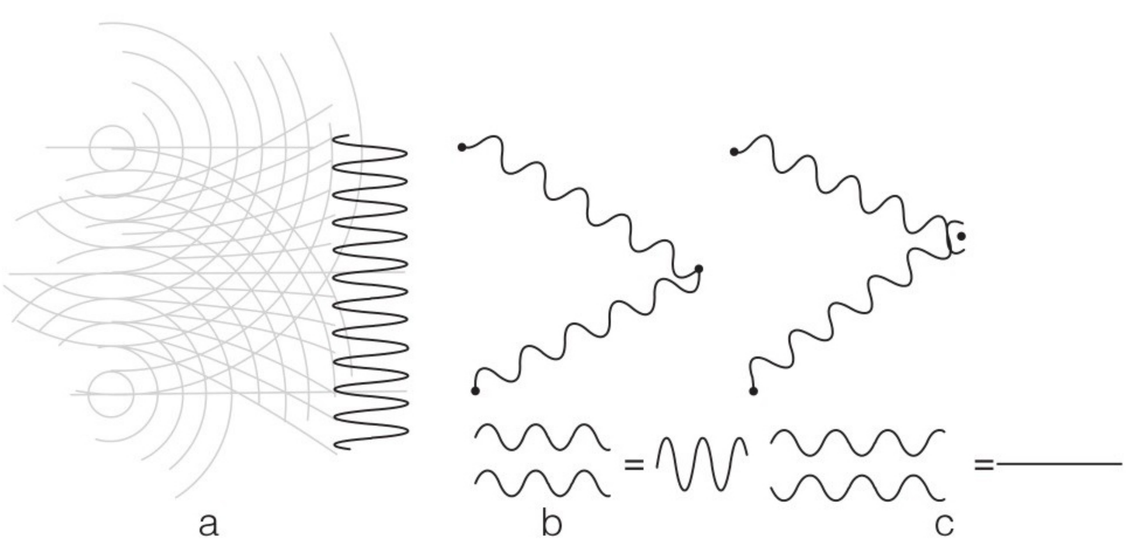


图17　a.水面上两个圆形波相互干涉，灰色的线是等相位线；b.两个波的波程相同，彼此之间发生相长干涉；c.两个波的波程相差半个波长，彼此之间发生相消干涉

这种现象被称为干涉，它是由两个波相遇时的振幅相加而产生的。如果两个波的峰值重合，则波峰的振幅变为原来的两倍。也就是说两个同相位的波，会发生相长干涉，即波的振幅加倍，见图17b。如果两个波是反相位的，也就是说其中一个波的波峰和另外一个波的波谷相遇，则合成波的振幅为零，使得这两个波彼此抵消，被称为相消干涉，见图17c。很显然，这样的现象不可能发生在粒子上，因为两个粒子是不可能互相抵消的。

1803年，托马斯·杨在一个著名实验中观察到了干涉现象，这一发现使得光的波动说成为解释光本质的主要理论。杨的实验简单而精妙。他用一个蜡烛当光源，蜡烛后面放一块屏幕，在屏幕上有两个距离很近的小孔。光线透过这两个小孔，投射在被放置在不远处的第二块屏幕上。如果只使用一个小孔(例如盖住另一个小孔)，那么在第二块屏幕上就会出现一个小小的光斑。然而，在两个小孔都开放的情况

下，奇妙的现象发生了：第二块屏幕上出现的并不是一个两倍于之前亮度的光斑，而是在此基础上出现了条纹。这些条纹由亮度几乎为零的直线组成，方向垂直于两个小孔的中心连线。图18是光线穿过两个小孔之后产生干涉现象的横截面图。这种条纹被称为“杨氏条纹”，是光作为波运动的关键证据之一。

我们之前提到牛顿做过的一个实验，光在两个靠得非常近的平面上反射时会产生彩色条纹。那么如何利用干涉原理来解释牛顿观察到的彩色条纹呢？我们知道，产生干涉现象需要两个波，其相对相位(两个波峰值的相对位置)可以调整。在牛顿的实验中，一束入射光被两个平面反射，从而被分成了两个波，就是在这两个波之间发生了干涉。如果两个反射面的距离等于光的波长，那么两个波的波峰相互重合，会形成一道亮条纹；如果两个反射面的距离等于半个波长，那么一个波的波峰和另一个波的波谷重合，从而发生相消干涉，产生“暗条纹”。因此，当你观察明暗条纹时，你会发现明暗条纹之间的间隔小于一个波长。对于波长约为500纳米的绿光，这一间隔甚至可以小于250纳米——约为头发丝直径的1/40。

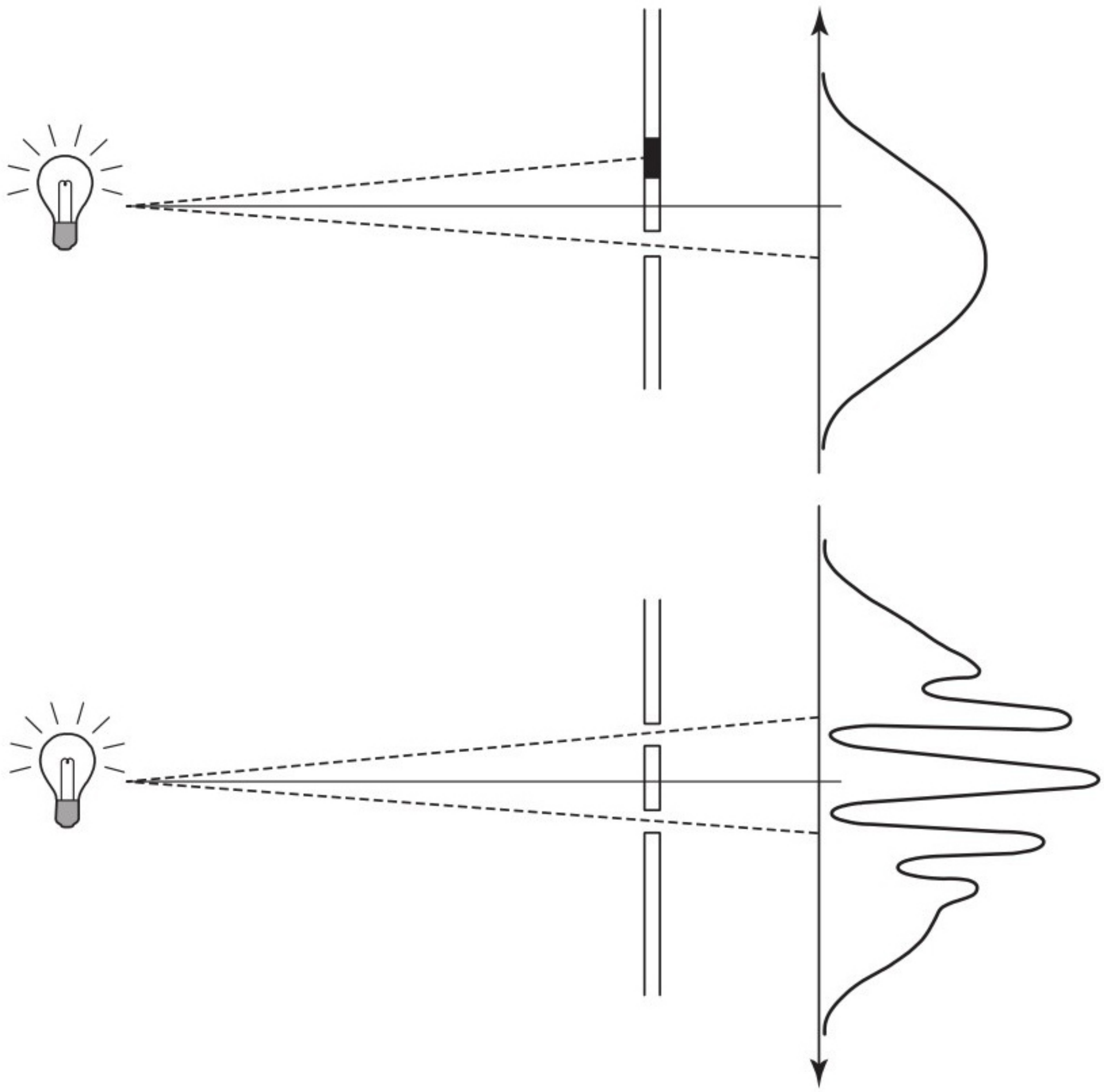


图18　托马斯·杨的实验。光透过一个小缝之后形成一个光滑的光强分布图。光透过两个小缝之后，形成一系列明暗条纹，这是典型的波动特征

当然，对于不同的波长，明暗条纹会出现在不同的地方。由于白光是由各种波长不同的光组成的，所以如果入射到平面上的光是白光，那么出现的条纹就会是彩色的。水面上油层边缘出现的彩色轮廓就是由于光波的干涉产生的。

干涉可以将微小的距离(与光的波长在同一个量级)转化为非常明显的光强度变化：最暗的地方光强度可以为零，而最亮的地方，其亮度可以达到单束光强度的4倍。这种光强度的变化很容易被探测或观察到。因此，在测量光波长这一量级的位移尺度时，干涉是一个很好的测量方法。很多光学传感器都是基于干涉效应的。

全息技术

干涉还可以用来制作真实的三维图像，它可以从任意角度观察图像，显示观察对象不同角度的图像。这种图像被称为全息图，与3D电影中所谓的合成图像不一样。全息图是通过记录物体散射出的光的完整波形制作而成的。我们平常拍摄的二维图像只编码了波的强度，而波的相位信息却丢失了。这是因为拍摄二维图像所使用的传感器只对波的强度做出响应，所以我们无法从这些图像中提取出相位信息。然而，要想对物体的形状进行编码，就需要利用相位信息。

干涉可以将相位信息编码为强度信息，这样光电探测器就可以记录目标波完整的振幅和相位信息，原理如图19所示。物体散射的光波与一束参考波发生干涉，其中参考波是由激光产生的已知形状的波。干涉图案则由传感器或者感光材料记录下来。这就是丹尼斯·加伯(Denis Gabor)于20世纪中期发明的全息技术。

与普通照片相比，观看全息图要复杂一些。首先，用一束参考波照亮全息图，其中一些光从全息图的编码图案中散射出来。这些散射光束有一个显著特性，它们再现了从原始物体散射出来的光束，因此当你的眼睛接收到这些散射光时，看起来就好像原始物体在你面前被重建了。在全息图周围移动时可以看到物体的不同侧面，因为这些从不同部分散射的光束编码了不同的信息。

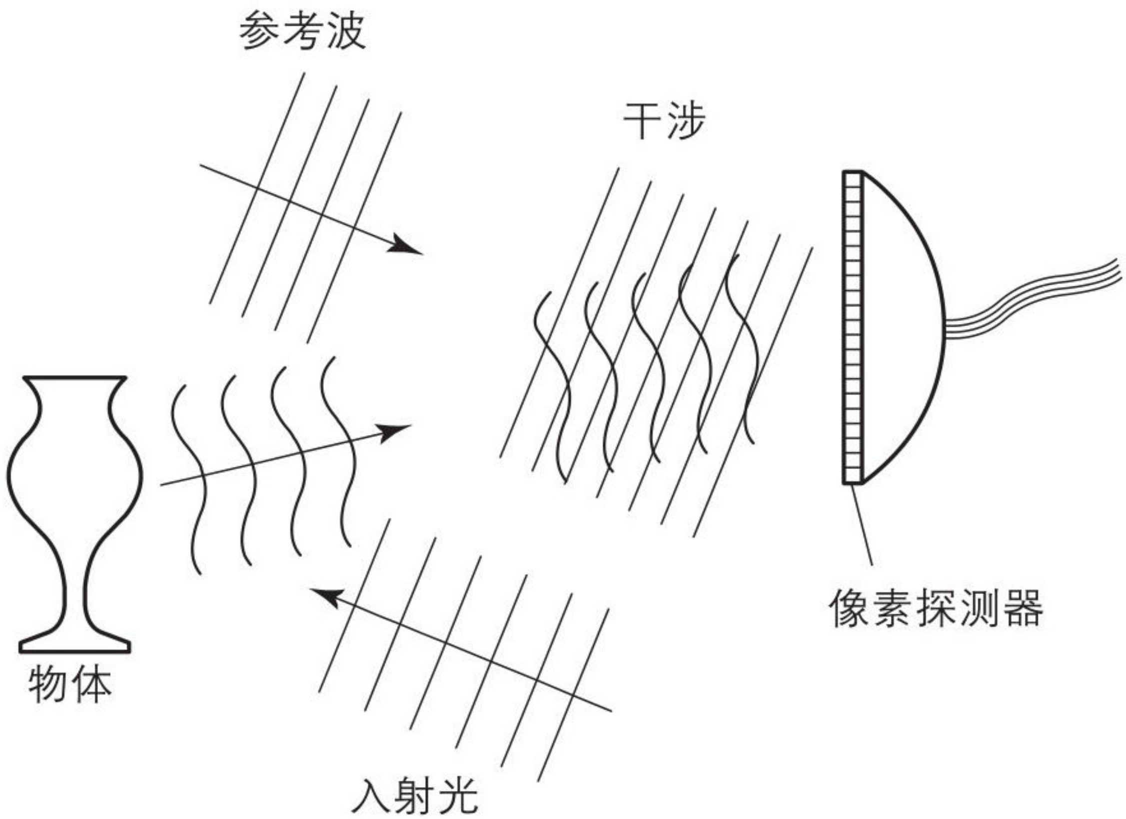


图19　全息图是通过记录一束参考波和物体的散射光波之间的干涉条纹而形成的

全息图也可以由电脑制作并压印在金属或其他材料上。材料表面的起伏模仿了参考波与物体散射波的干涉图样：凸起的部分代表了亮条纹，凹进去的部分代表了暗条纹。同理，想观看这样的全息图也要用一个参考波照射，使得其材料上散射的光再现原物体发出的散射波前。这种全息图还被用来作为安全装置，包括在钞票上的使用[1]，因为它太难制作了，要借助很先进的工程技术才可以。

再次探讨成像极限

光的波动说还解释了为什么我们无法用显微镜观察极其微小的物体，正如阿贝所注意到的一样。小到半个微米(一米的百万分之一或可

见光波长的一半)的微小物体可以用一般的光学显微镜观察，而对于再小一些的物体，我们就需要用更加复杂的方法进行观察，这是因为光的波动特性限制了光斑的最小尺寸。

我之前提到过，两束光相遇会发生干涉从而产生暗条纹，也就是强度为零的区域。这些条纹的间距取决于两束光相交时的角度。如果角度很大，则条纹间距较小；反之，如果角度较小，则条纹间距较大。条纹的最小可能间距为一个波长；对于可见光而言，这个间距大约是一个微米。

如果这个干涉条纹图样被记录为全息图，那么当它再次被参考光束照亮时将会产生两束光，其方向与用来记录干涉条纹的光束方向一致。如果想用显微镜观察到这种条纹图，使用的透镜必须将这两束光都捕捉到才能形成干涉条纹，如果透镜只能捕捉到其中一束光，那么观察到的图像中就不会出现干涉条纹了。

这是我在第2章中介绍的阿贝准则的物理基础：成像系统的透镜捕捉到的两个光束之间的最大角度决定了所能观察到的物体的最小尺寸。很容易看出，透镜系统所能观察到的物体的最小尺寸大约等于穿过透镜的入射光的波长。因此传统的光学显微镜能够观察到的最小物体尺寸约为人的头发丝直径的1/50，比这再小的尺寸就观察不到了。例如，光学显微镜可以用来观察生物细胞，但不能用来观察细胞核。

超分辨率成像

光学科学家和工程师们想出了很多巧妙的方法，以突破传统光学显微镜对观察目标尺寸的限制，这样他们就可以看到细胞内部，或者可以观察到尺寸不到光的波长百分之一的物体。这些仪器使用了新材料和新工艺，比如把纳米级粒子附着到观测目标上，或者将会发光的分子插入细胞中。当它们被一束短波长的光照射时，可以发出长波的

光(如荧光)。由于它们的尺寸比显微透镜的分辨率小得多，根据阿贝公式，最终得到的图像会是一个尺寸受限于显微镜光学的斑点。但是我们可以用摄像机长时间观察附着在物体表面的纳米颗粒发出的荧光，并且确定光斑强度最强的位置，从而精确定位图像的中心点。这个技术被称为“光激活定位显微镜”(photo-activated localization microscopy，缩写为PALM)，由美国的埃里克·白兹格(Eric Betzig)发明。这个发明彻底改革了活细胞成像技术，使得人们可以在宽视野范围内更快地采集信息并获取更精确的深度分辨率。

在较大尺寸的荧光物体中测量微小结构的另一个方法是，先用一束光照射物体使其产生荧光，然后用第二束圆形光照射物体，使得光斑外围的荧光消失，只有中心的荧光点保持不灭。通过这样的方法，可以利用保留下来的尺寸很小的荧光点对物体进行精确定位，精确定位的方法跟我们之前描述的定位方法一样。这种方法叫做“受激发射损耗显微技术”(Stimulated Emission Depletion Microscopy，缩写为STED)，是由德国的斯特凡·赫尔(Stefan Hell)发明的。在第5章我会对受激发射的过程进行详细描述。这些高分辨率成像的新技术使得科学家们得以观察细胞内部的结构，从而在生物学和医学领域中产生了巨大的影响。这种影响的重要性已经获得了认可——白兹格和赫尔获得2014年的诺贝尔化学奖。

阿贝准则反过来使用也是成立的：当光通过显微透镜照射在观测样本上时，聚焦形成的光斑直径不可能小于一个波长。而且聚焦的紧密度，也就是光斑的大小，取决于透镜和透镜照射面之间的夹角范围：角度范围越大，光束聚焦得越紧密。

干涉光束的角度范围和明暗条纹尺寸之间的关系是波的一个基本特性。19世纪初期法国科学家约瑟夫·傅里叶(Joseph Fourier)对这一观点进行了量化，对光波的传播进行了详细的数学分析。傅里叶定

理简单来说就是：要想使光线聚焦的光斑尺寸越小，那么就得保证传播到这个光斑的入射光角度范围越广。

衍射

这解释了光的另外一个特征，即光在传播过程中会逐渐发散。这是因为根据定义，一束光的空间范围是有限的，它必须由一道道沿着不止一个方向传播的波组成。我们可以用激光笔来验证上述想法。激光笔发射出的光束直径大约是10微米，当它照射在屏幕上时，直径大约为1毫米。如果激光照射的距离更远，例如照射到月球上去(大约40万千米)，那么光斑直径会高达24千米。这种现象就是衍射。

衍射在测定结构的形状和对称性方面有一些有趣的应用。例如，一束光照射在有小孔的屏幕上，当小孔的直径和波长相近时，光线会通过小孔发生衍射，且其光束的扩散程度与孔径的尺寸成反比。这些衍射光束在距离屏幕一段距离的地方会相互发生干涉，形成干涉条纹，也就是所谓的衍射图案，它反映了小孔的大小及相对位置。例如，如果小孔按照规律进行排列，那么经过该小孔产生的衍射图案就会显示出相应的规律性。使用衍射图案对物体进行测量的优势在于不需要非常昂贵或者复杂的透镜系统，也不需要让探测器非常靠近物体，只需要观察因为衍射而自然扩大的图案。

现在，我们假设屏幕被一个透明的固体材料所替代，比如说晶体蛋白结构。“小孔”则被蛋白质分子中的原子所替代。这些原子非常小，并且通过分子中的键相互连接，这些键的长度约为十亿分之一米(0.1纳米)。如果波长接近于这个尺寸的光照射在这种结构上，光就会发生衍射，分子本身的实际结构就可以由衍射图样确定出来，这是**X射线衍射**的基础。正如第1章中提到的，它曾因探索DNA的结构而闻名，现在也是生物化学领域很常见的工具，常常被用来探寻新分子(例如可能有助于开发药物的新分子)的结构。这一过程需要一束明亮的X

射线源，以及将分子晶体化的方法。图20是牛肠道病毒[2]晶体的衍射图。

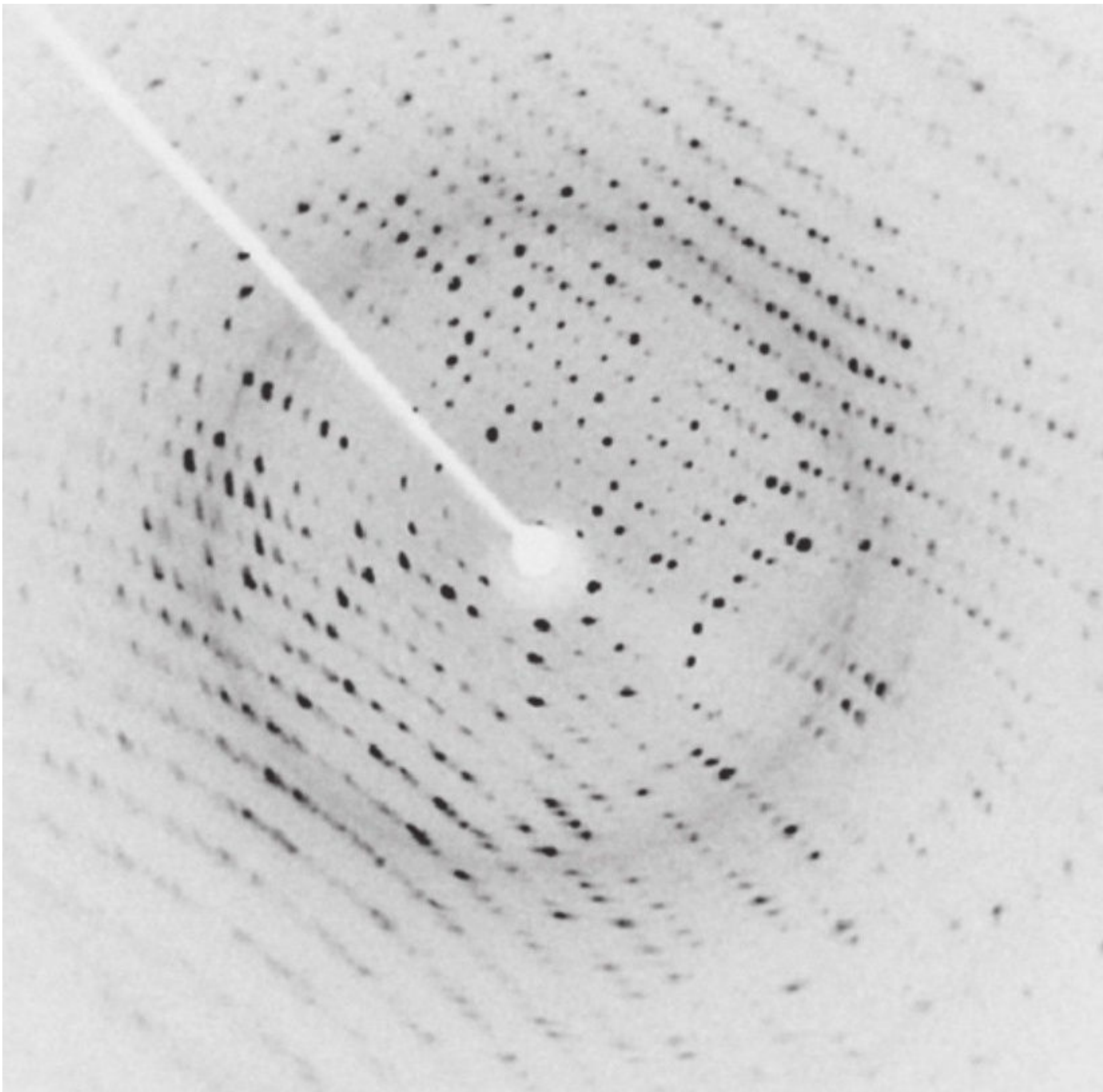


图20　使用现代同步辐射X射线源拍摄到的蛋白质晶体的X射线衍射图

很明显，如果你想远距离传输光，那么衍射可能会是一个问题。衍射会使得光束的能量分散开来，因此，随着传输距离的增加，你需要的光学系统和传感器要越来越大才能够接收所有的能量。这对电信

业来说非常重要，因为几乎所有通过远程通信传输的信息都被编码在光束里。

导波

为使远距离通信成为可能，就要对衍射进行管控，解决这一问题的方法就是使用波导，例如光纤。波导是一种经过精心设计的截面具有特殊折射率的结构。例如，光纤内的折射率变化分布经过了特殊设计，使得直径为几百万分之一米的“芯”比周围“包层”的折射率更高。这样，光就被局限在折射率较高的“芯”里(成为导波)，沿着光纤移动而不发生衍射，从而完成远距离传输(例如通过海底光缆横跨大西洋)，同时光束的大小保持不变。从通信到传感器等许多光信息基础设施，都是依靠这种控制光的方式来工作的。

偏振

光的波动模型的最后一个重要特征是偏振特性。回想一下，在横波中，波的振动方向与波的传播方向相垂直。但是，与波的传播方向成直角的方向有两个，也就是说，横波有两个可能的波动方向。

以一根绳子上产生的波为例。如果你将绳子的一端上下快速移动，你会发现波是沿着绳子移动的。如果快速地左右移动绳子一端，也会发生类似的事情。不管是垂直于地面振动，还是水平于地面振动，这两种振动方向都与绳子的方向垂直，也就是与波的移动方向垂直。这样的波都被称为“横波”。

光波就是一种横波。水平偏振光有一个在水平面(相对于光具座[3])上振荡的电场。类似地，垂直偏振的光束在垂直平面上振荡——还有其他更加复杂的偏振形式，这里就不赘述了。要解释双折射现象，我们可以先从晶体的结构入手。晶体由单位晶胞重复排列组合而

成。作为构成晶体的最小几何单元，晶胞是由原子构成的特殊结构。这些晶胞本身可能是不对称的，光沿着晶胞的长轴或短轴传播时会发生不同的偏振，折射率也将随之不同。因此，当光通过这种晶体的时候，光的传播方向会发生不同程度的偏移。

众所周知，太阳眼镜就是利用了光的偏振特性。一些太阳眼镜使用塑料(例如塑料偏光薄片)做镜片来充当偏振片。偏振片只允许某个特定偏振方向(例如垂直偏振)的光通过，垂直于该特定方向的偏振光(例如水平偏振)则被偏振片所吸收。塑料偏光薄片是由橄榄球形状的分子组成的，这些分子在这种塑料聚合物中整齐排列并且像被“冻住”一样一动不动。这些分子优先吸收沿着分子排列轴方向偏振的光。一般来说，由于由物体散射的阳光会随机偏振(约为每个偏振方向的50%)，因此，过滤掉某一偏振方向的光就可以有效地将场景亮度降低一半。此外，太阳镜还可以减少眩光，即从光亮的平面(例如汽车的引擎盖或者风挡玻璃)反射的光线。这些平面倾向反射偏振方向与该平面平行的光[4]。戴上我们上面所说的太阳镜之后，这种反射光就因为其偏振方向被太阳镜挡住，因此看马路的时候视野会更加清晰。

透明双折射材料也可以在不吸收光的情况下改变光的偏振。这是因为光的传播速度取决于相对于材料“方向”的光的偏振方向。一些材料，例如一般的玻璃，是没有特定方向的：你可以任意旋转它而不改变其对光束的影响。但如前文所述，双折射材料中原子的排列有一个优先方向，即对称轴的方向。沿着这个方向，原子对光的响应是不同的。也就是说，沿着对称轴偏振的光的传播速度比垂直于对称轴偏振的光更慢。想象一下，光的偏振方向与对称轴的夹角为45°时，假设有一半的光沿着对称轴方向偏振，另一半则垂直于对称轴方向偏振。如果后者的速度减慢得足够多，穿过材料之后透出来的光，将会沿-45°的方向偏振，相当于光的偏振方向被“旋转”了90°。

一些双折射材料可以通过主动调整分子排列轴的方向来控制偏振状态，例如对材料本身施加电压。一个典型的例子就是液晶(Liquid Crystals，缩写为LCs)，它由细长的分子组成，液晶中分子的方向可以通过施加电压的方式来控制。此外，通过施加压力或者应力，其他的一些材料也可以变成双折射材料，这是因为外力使得材料内部的分子发生“转动”或者让原子的排列方向发生了改变。利用这个现象可以构造力学传感器，通过观察光传感器端输出的光的偏振状态来对力进行监测。

在两个偏振片之间放置一个双折射液晶，就可以通过电对光的强度进行控制。施加电压可以使分子重新定向，从而改变偏振光束在这一材料中的折射率。如果在液晶之后放置一个偏振片，那么根据施加的电压的高低可以控制通过偏振片光强的高低。将这样的“单元”构成阵列，且每一个单元由独立的电信号驱动，这样就构成了一个显示屏，每一个单元就是一个像素。这就是液晶显示器(Liquid Crystal Display，缩写为LCD)的基础，这一技术通常用于电脑显示屏或者电视机。

事实上，这种显示屏还可以用来播放3D电影。在这样的电影中，我们感知到的深度其实是一种错觉，来自人类视觉的立体感。由于我们的两只眼睛间有几个厘米的距离，所以两只眼睛观看同一个场景时感知的方位略有不同。这两幅图像在大脑中结合，让我们感知到深度。

这种错觉可以通过3D眼镜的偏振作用再现出来。两幅图片被同时投影在显示器或屏幕上，每一幅都是由特定偏振的光产生的，而且是从略微不同的角度进行拍摄的。3D眼镜由偏振方向不同的两个偏振片组成，使得左眼可以看到其中一幅图，右眼可以看到另外一幅。因为两幅图片发出的光都分别只与其中一个偏振片允许通过的方向一致，

因此每只眼睛只能看到一幅图片，于是我们看到的场景就跟我们在自然界看到的景象一样，即物体和人看起来都是三维的。

光的波动模型的成功令人振奋，这让我们得以了解光的一些重要特性，从而利用这些认识来构建新的技术。光的粒子说同样令人惊叹。然而，这两种截然不同的关于光的本质的学说都是必要的，这确实令人十分困惑。我现在就要转向这个难题。

[1] 例如20英镑的纸钞上就印有18世纪苏格兰经济学家亚当·斯密(Adam Smith)的全息图。

[2] 一种小RNA病毒，下图为其衣壳蛋白结构。

[3] 一种多功能的通用光学仪器。

[4] 这个现象于19世纪被大卫·布鲁斯特(David Brewster)爵士发现，并以他的名字命名。

04 光的波粒二象性 Duality

有两种看待光的不同观点，一种将光视作粒子，另一种则将光视为波，它们都包含着深刻的洞察力与价值。这两种不同的观点都给人们带来了启发，不仅帮助我们加深了对自然世界的理解，还使相关新技术的开发和设计成为可能。然而值得注意的是，这两种观点中对光究竟是什么的理解似乎大相径庭。一方面，粒子模型将光视为一个固态的实体，具有能量，并沿着明确的轨迹移动；另一方面，波动模型将光描述为一个可扩散的实体，它穿过空间，与固态物体的运动无关。这两种观点怎么可能描述的是同一个概念？惠更斯及其同时代的人早就认识到了这种困境。作为对光本质的解释，这两种不同的观点针锋相对，一直持续到19世纪。

当麦克斯韦发展他的电磁场理论时，他发现能够用此来解释光的波动性，正如我们在第3章中看到的那样。这一推理的胜利似乎证实了托马斯·杨和奥古斯丁·菲涅耳分别做的两个实验(见第3章)，在这两个实验中他们揭示了两个基本现象——干涉与衍射，它们都不适宜用粒子模型解释。然而，光作为粒子沿着轨迹运行的概念仍然非常强大，它可以用来分析和设计光学系统。因此，科学家们需要重新考虑一下，有没有方法使光的粒子说与光的波动说握手言和呢？

再次审视光的轨迹

17世纪上半叶，法国人皮埃尔·德·费马就折射现象提出了一种巧妙的解释，与斯涅尔的解释截然不同。让我们回顾一下斯涅尔定律，它描述了光在两个透明介质之间的界面处发生折射，即光的传播方向发生变化。其中，光用相对于界面的入射角度和撞击界面的位置来描述，传播方向的改变程度与这两种透明材料的折射率之比成正比。因此，最重要的似乎只是光和界面的局部特性。斯涅尔定律适用

于轨迹上的任意一个点，就好像光可以凭着自己的“感觉”，在遇到新界面时调整方向。

费马的想法则完全不同，他认为应该根据起点和终点来定义光的轨迹，如图21所示。他认为最应该问的问题是：光穿过空间中两点之间的轨迹是怎样的？在他看来，这个路径应该是耗时最短的路线。基于这一想法，费马异曲同工地给出了与斯涅尔相同的答案。这一成果是非凡而深刻的，因为费马的“最短时间原理”告诉我们，当我们想到光的时候，要考虑整体情况：入射方向、初始位置、最终位置以及处于两个位置之间的一切因素。这一理论与光的粒子模型(粒子对其直接接触的周围环境进行反应的局部模型)相比，差别不言而喻。

这个想法被德国自然哲学家戈特弗里德·威廉·冯·莱布尼茨(Gottfried Wilhelm von Leibniz)所接受，他是与牛顿同时代的科学家，也是牛顿的竞争对手。莱布尼茨对费马所描述的折射过程的整体图景，以及其中所隐含的“优化”概念印象深刻：光对整个空间进行探索，最终仅选择在指定起点和终点之间传输时间最少的路径。于是他开发了数学工具来分析这个想法，通过计算变化的微积分，可以计算出运动轨迹的微小变化对传输时间产生的影响。莱布尼茨认识到了费马原理中所提出观点的重要性，即光通过从一点到另一点的运动定义了“最佳”轨迹。

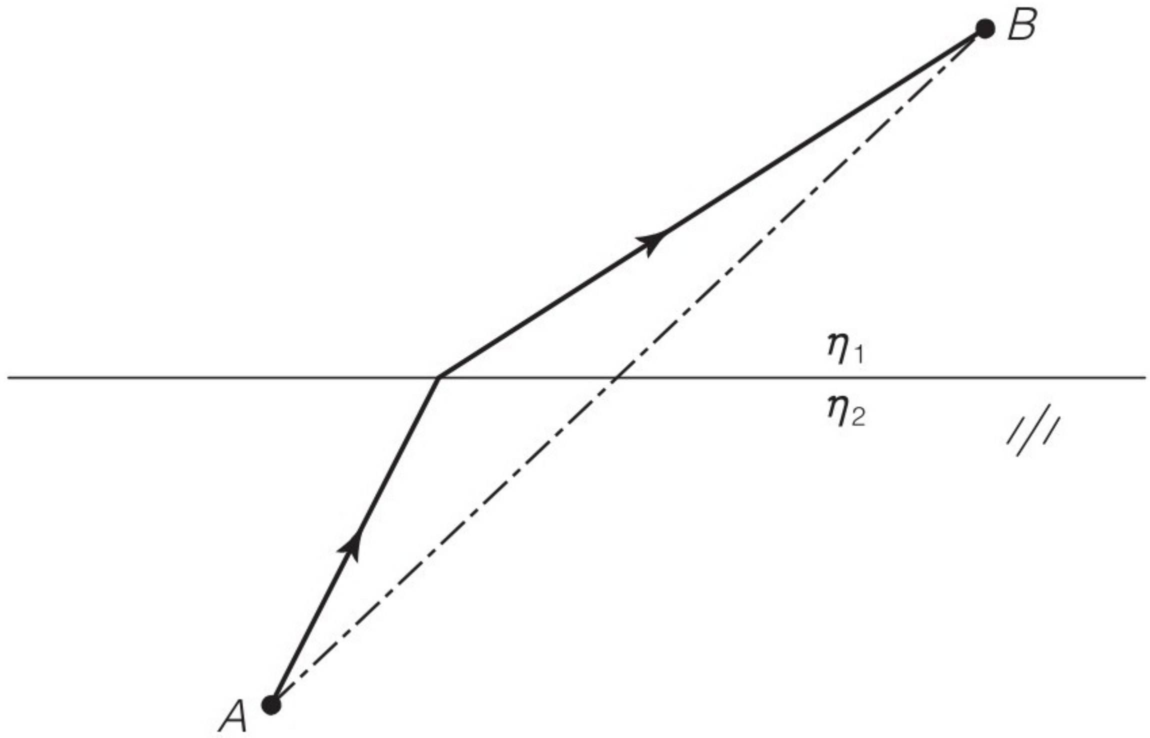


图21　费马认为光传播的路径是连接起始点(A)和终点(B)且传播时间最短的路径。光穿过两个光学介质之间的界面，且光在两种介质中的传播速度不同

事实上，莱布尼茨也被这种最优化的概念所吸引，他将它提升为一个目的论原则，即世界的方方面面都处在某一个起点和某一个终点之间的最佳轨道上。当把这一原则应用到科学领域之外时，这种立场便显现出了其固有的矛盾。伏尔泰(Voltaire)在他的小说《老实人》中便巧妙地讽刺了这一点，他将莱布尼茨的思想借潘葛洛斯博士之口说了出来。在书中，潘葛洛斯博士坚持认为任何发生的事情已经是穷尽所有可能性之后的最好安排，连自然灾害和人为灾难也皆是如此。

连接光波与光线

尽管如此，莱布尼茨的这一数学思想仍被证明是卓有成效的。19世纪著名的爱尔兰数学家威廉·卢云·哈密顿(William Rowan

Hamilton)开始研究这些问题。他展示了如何将光的波动概念与粒子概念相结合。波可以通过波长、振幅和相位来定义；粒子由粒子的位置、移动方向定义，而粒子的集合则由其密度(在给定位置的粒子数量)和方向范围来定义。光在介质之间传播时，介质的光学特征由它们的折射率决定，这可能因空间而异。例如，在图21所示的介质界面处，折射率存在一个阶跃变化。

哈密顿表明，光到底是表现出粒子性更多一些，还是波动性更多一些，主要取决于空间中折射率变化的速率与光的波长之间的关系。换句话说，如果折射率变化速度的数值非常接近入射光一个波长的大小，那么光的波动性特征将会非常明显；如果折射率的变化速度基本不变或者非常慢，那么光的粒子特性就会比较明显。

哈密顿展示了在某些常见情况下，较简单的射线图是如何从较复杂的波动图表现出来的。当光的波长与其传播介质的大小相当时，光会表现出明显的波动特性，从而出现如衍射或干涉的波动现象。因此，当光照射的物体直径仅为几微米或者具有非常锋利的边缘时，例如鸟羽或蝴蝶翅膀上的精细结构，你会看到衍射图样。在另一种情况下，例如相机的镜头，其透镜的折射率是均匀的，即折射率的变化速率为零，光会表现出明显的粒子性，因而可以用粒子的运动轨迹来解释。

此外，哈密顿还表明，费马提出的光的轨迹理论与一个波的特性直接相关，这个特性就是波前。当波在空间中传播时，将相位相同的位置连接起来就是波前。例如，一块石头被投入池塘中，你看到池塘表面的每一个涟漪，即圆形的图案就是波前。这些涟漪是池塘表面上的水波“达到峰值”(或低谷)时的波前。因此，哈密顿指出光线可以被认为是以直角与波前相交的线，如图22所示，从而将相邻的波前与一个定义明确的光的轨迹这两个概念连接了起来。

哈密顿的“光学类比”

这个显著的结果引申出了另一个深刻的类比，即哈密顿的“光学类比”。他注意到，在那些众所周知的力学公式中，固态物体的运动和位置均是基于轨迹的概念。那么，这样的轨迹是否像光的轨迹一样存在着某种“最佳”情况呢？18世纪的皮埃尔·路易·莫佩尔蒂(Pierre Louis Maupertuis)就曾对此展开过研究。

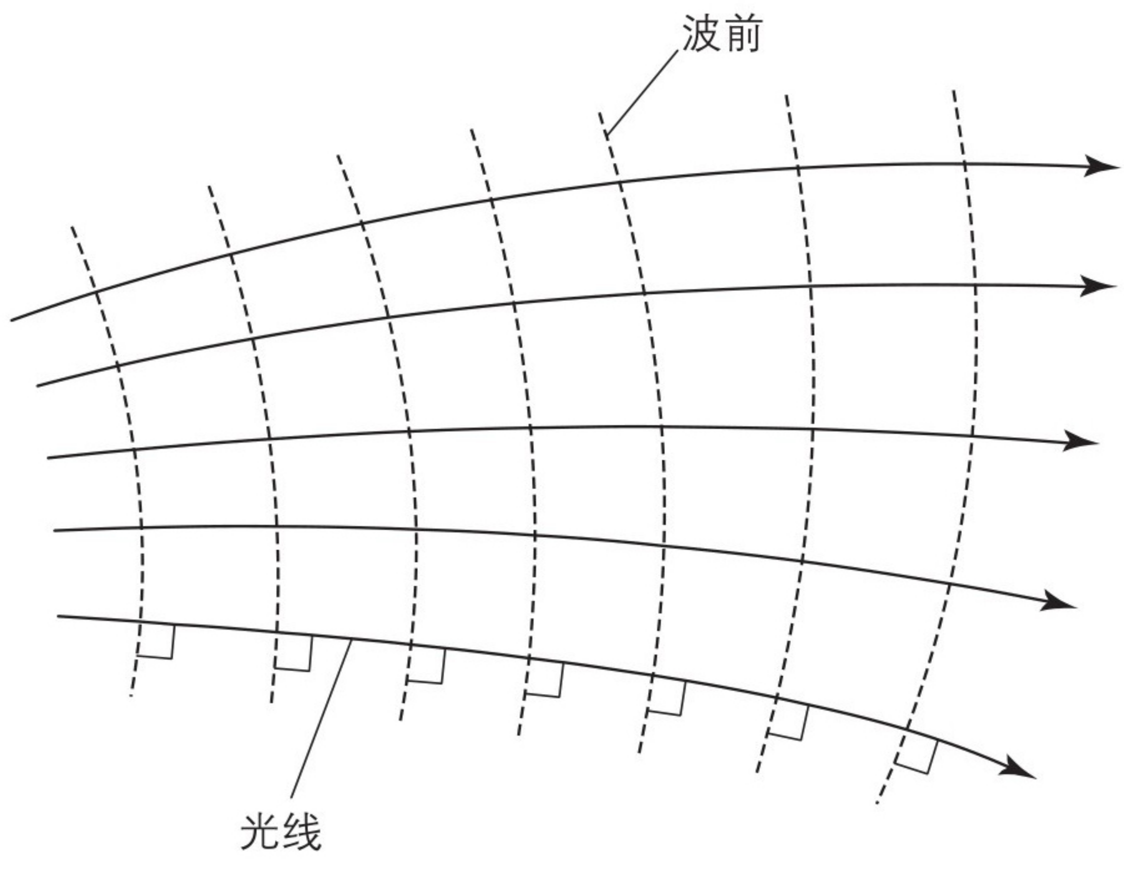


图22　哈密顿把光线与波前的概念联系起来，从而使光的粒子性与波动性结合起来

莫佩尔蒂制订了一种方法来评估“作用量”的最佳值，其中作用量指向物体的运动轨迹，是物体的速度、移动距离与质量的乘积。他认为，对于物体在两点之间运动的实际轨迹，作用量应该是最小的。

莫佩尔蒂的“最小作用量原理”在概念上与费马的“最短时间原则”非常相似。事实上，18世纪的瑞士数学家莱昂哈德·欧拉(Leonhard Euler)利用莱布尼茨的微积分，从莫佩尔蒂的原理推导出了著名的牛顿运动方程。由此，欧拉将“粒子通过环境感知其运动轨迹”与“粒子的运动路径受指定起点和终点之间的整个空间影响”这两种理论连接了起来。

哈密顿推导出了可以根据运动物体所处环境来描述物体作用量变化的方程。并且，这个方程与哈密顿为描述光线轨迹所推导的方程有着非常相似的形式(只不过在这一情况下，运动物体所处环境指的就是折射率如何随介质中位置的变化而变化)。因此，在固态物体的轨迹和虚构的波前之间存在着一种潜在的类比：也许所有物体都可能具有类似粒子的轨迹和类似波动的特性？事实上，哈密顿方程式及其同名函数对于思考理解光的下一个重大课题——量子力学来说非常重要。

未解之谜

除此之外，还有很多现象暗示着科学界仍然存在许多新机遇，等待着人们去揭开谜底。大约在19世纪末期，即便哈密顿已经将光的粒子性与波动性联系起来了，仍然有一些关于光的未解之谜，这些谜题用主流模型是无法解释的。其中最重要的两个未解之谜，一个是有关热物体(包括太阳)的颜色，另一个则是有关不同原子在火焰中的颜色。

当物体被加热而升温时，它的颜色会发生改变。以一块金属为例，随着温度的升高，它首先会发出红色的光，然后是橙色的光，接着则是白色的光。这一现象背后的原理是什么呢？这个问题困扰着当时许多伟大的科学家，包括麦克斯韦本人。依据麦克斯韦的理论，随着温度的升高，物体发出光的颜色理应变得越来越蓝，并最终发出不

处于人类视觉范围内的紫外线。显然，这与现实生活中观察到的现象大相径庭。

第二个未解之谜则集中在研究原子发出的光上。约翰·巴耳末(Johannes Balmer)在这一问题上作出了开创性的贡献，我们将在第5章中更详细地研究这种机制。光谱中的颜色分布是光的重要特征，从这个角度来说，原子发出的光与太阳光非常不同(太阳是热物体的一个很好的例子)。太阳光具有我们非常熟悉的“彩虹”光谱(见图23a)，由从红色到紫色间的所有连续的颜色组成。相比之下，一组原子发出的则是一组离散的颜色(见图23b)，即一组特定波长的“光谱线”。这些光谱线与所涉及的特定原子的内部结构有关。

这两种现象要求我们彻底修改对光的理解，因为它们无法用波动或粒子的当代模型解释。

19世纪后期在柏林洪堡大学工作的马克斯·普朗克(Max Planck)首先提出了一个想法，用来解释热物体发出的光谱，这种热物体也常被称为“黑体”。他推测，当光和物质相互作用时，它们只能通过交换离散的“小包裹”来实现，这种“小包裹”可以是量子或能量。普朗克认识到他的想法是非常具有颠覆性的，尽管这将极大地改变我们对光的看法，但他仍不愿意太多地从这个角度去推断有关光的本质。他的想法使光重新被看作一种粒子，一种离散的带有固定能量的物质，可以被原子吸收或发射。

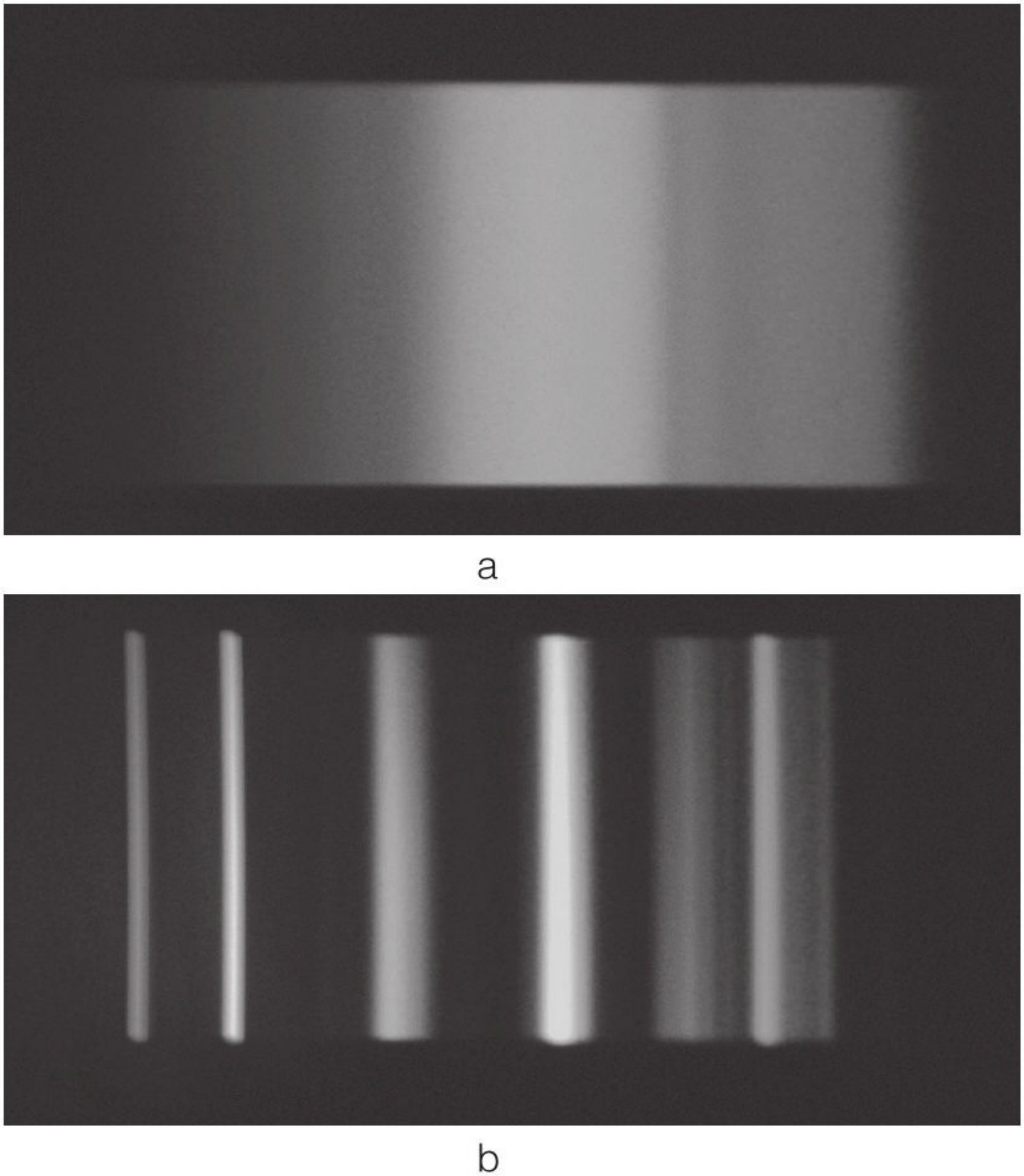


图23　a、b分别是太阳（一个“黑体”）和霓虹灯发出的光谱图。前者有连续的色带，后者则显示出特定颜色的离散谱线，这些谱线是氖原子的“指纹”

科学家们对光的理解似乎发生了倒退。毕竟，光的波动模型已经解释了迄今为止观察到的所有现象。并且，从哈密顿的工作中可以清

楚地看出，即使是光的粒子性表现得最明显的现象，如光沿着特定路径传播，也可以用光的波动模型来解释。所以，光由粒子构成的这种想法似乎没有必要进行讨论了。当然，把光重新看做粒子只是一个计算性的“修复”，来解释现有理论无法解释的现象，它最终会被一个更合理的理论所替代。然而，结合巴耳末的观察实验，这个猜想终将从根本上改变物理学。

在普朗克提出光和物质之间进行离散能量交换的想法之后的几年里，阿尔伯特·爱因斯坦利用这一想法解释了另一个长期无解的物理现象——光电效应。当光照在金属上时，一些电荷(带电的电子)会从金属中喷射出来，这就是光电效应。电子射出的速度取决于光的波长。光必须足够“蓝”，即具有足够短的波长，才能够使电子喷射出来。随着它变得越来越蓝，电子以越来越高的能量射出，速度也越来越快。

爱因斯坦指出，电子至少需要具有某一特定的能量才能脱离金属的束缚。不仅如此，他认为光的粒子中具有离散的能量，且这样的能量与光的频率成比例(其比例常数被称为普朗克常数)。因此，当光照射到金属表面时，如果光的频率足够高(波长足够短)时，光的粒子可以将能量传递给电子，为电子提供足够的能量从而逃离金属的束缚。爱因斯坦的模型表明，光与物质之间离散能量交换的起源来自光实际的离散特征。这标志着光的粒子模型的完全复兴。

这一想法与巴耳末对原子产生离散光谱线的观察结果非常吻合。但是，为了完整解释离散光谱线的现象，显然需要着重解释为什么原子会通过这种“小包裹”的能量去发射光。彼时在曼彻斯特工作的丹麦物理学家尼尔斯·玻尔(Niels Bohr)提出了解释这一问题的关键。他认为，光之所以作为离散能量的“小包裹”被发射，是因为原子本身只能以某种构型存在。他认为原子类似于微小的行星系统：电子在

围绕中心核的轨道上运转。电子可以在两个稳定的轨道之间“跳跃”，同时发射或吸收光。究竟是发射还是吸收光，则取决于电子是跳跃到更低还是更高能量的轨道上去。这些轨道或量子态的特征取决于原子本身：该原子有多少电子，还有其原子核的大小。因此，电子在两个量子态之间移动并且发射或吸收能量，是由原子本身的特性所决定的。这就是说，当光子的能量与原子中电子的两个量子态的能量差相同时，光的吸收或发射就成为可能。玻尔的想法巧妙地解释了巴耳末的观测结果，并肯定了光束作为离散粒子集合的想法。

所有这些进展都有可能破坏麦克斯韦理论所强烈肯定的光的波动模型。它们甚至超越了哈密顿试图将光的波动性与粒子性相协调的努力，因为这一全新的想法，即光束是离散粒子的集合，似乎是光的基本特质，而不仅是拿物体的大小与光的波长进行比较的结果。因此，科学家们重新审视了光的本质问题。

1908年，在剑桥工作的杰弗里·泰勒[1](Geoffrey Taylor)用极其微弱的光进行了杨氏双缝实验。光非常微弱，以至于任何时刻同时通过两个狭缝的光子平均下来不到一个，但他仍然看到了干涉条纹。这个结果很奇怪。如果我们认为从光源到探测器有两条路径，第一条路径是通过第一个狭缝，第二条则是通过另一条狭缝，那么一个光子从光源到达探测器的路径也有两条。然而，在同一时间内通过双缝的光子数只有一个，那么，单个光子是怎样形成干涉条纹的呢？这让当时的科学家们陷入了两难的境地。玻尔解决了这一难题。玻尔指出，一方面，光子总是会选择这两种路径中的一种，而另一方面，它表现得好像它同时通过了这两种路径一样。因此，即使是单个粒子也可能表现出类似波的行为。

光的波粒二象性

正如你可能想象的那样，要摆脱这个难题，需要一个真正具有颠覆性的想法。20世纪20年代在剑桥工作的物理学家保罗·狄拉克认为，光的基本属性在于它既是粒子又是波，两种特性同时存在。现在，这一想法对你来说可能只是诡辩，一种没有回答任何问题的逻辑游戏，但它背后隐藏着深刻的洞见。狄拉克发展了麦克斯韦电磁场理论的量子力学版本。利用这一理论，狄拉克能向你展示，如果你使用像杨氏双缝干涉仪这样的装置来测量这些“量子场”，你会看到干涉效应这种体现光的波动性的现象。然而，如果你只是测量光的强度，那么只用去数光束中的光子数就可以了。

这一理论非常深刻，它将量子场设定为构建宇宙的基本实体——它不是粒子也不是波，而是既是粒子又是波，具有完全的波粒二象性。它完美地解释了光所展现出的所有现象，并提供了理解所有光学效应的框架。这些光学效应不仅包括牛顿、麦克斯韦和哈密顿的经典世界，还包括了普朗克、爱因斯坦和玻尔的量子世界。但是，这一理论实在是令人费解，因为它包含有一个完全非直觉的实体——量子场。光只是量子场的一个例子。

光既是波又是粒子这一跨时代的理论激发了一些重要的新思想。例如，路易斯·德布罗意(Louis de Broglie)提出，如果这种波粒二象性存在于光中，那肯定也应该存在于所有其他事物中。因此，那些我们通常认为的由粒子构成的物质实体也应该具有“波动”的特征。他的想法超越了哈密顿所考虑的范围，甚至找到了物质波动性波长的定义。这一波长现在被称为德布罗意波长λ_{dB}，它与粒子动量(质量与速度的乘积)成反比，比例常数为普朗克常数h。

$$\lambda_{dB} = h / mv$$

这个波长公式表明，如果想观察到这种波动现象，必须使用质量非常轻或温度非常低(意味着移动速度非常慢)的粒子。使用分子代替光通过双缝干涉仪，一样可以形成干涉图样，如图24所示。这个结果简直令人难以置信。如果你认为分子只是非常轻的粒子，那么你将无法解释这个干涉图样的出现，因为你认为这个粒子只能通过这两个缝隙中的一个。然而，一个具有质量的粒子竟然可以同时通过两个缝隙并发生干涉，这个想法是非常惊人的。



图24　一次只让一个分子通过一个微缩版的杨氏双缝装置所形成的干涉图样。这两个很小的狭缝之间的距离仅为十亿分之一米

埃尔温·薛定谔(Erwin Schrödinger)猜想，如果所有物质都表现出波的特性，那么肯定存在着一个描述其波动性的方程。从哪里开始寻找呢？他利用哈密顿通过“光学类比”推导出的方程，来描述粒子

的“作用量”是如何演化的。通过对这一方程进行一个简单的补充(用到了普朗克常数)，薛定谔塑造了一个描述波动的函数。这就是薛定谔著名的“波函数”的来历。波函数具有许多与光波非常类似的特性，包括干涉和衍射，但是其描述的对象仍然是前量子力学时代语境下的具有明显质量与重量的粒子。那么，波函数描述的到底是什么呢？是实际的粒子本身，还是我们对粒子认知的某种简化呢？

光的强度是光的一个重要性质。如果将光看作波，那么光的强度与其振幅的平方成正比。如果将光看作粒子的集合，那么光的强度则与光束中的光子密度直接相关。类似地，波函数的平方与特定时间与空间中处于特定点的粒子的密度有关。但是，要想确定粒子在特定时刻的空间位置是不可能的。这种不确定性似乎是世界的基本属性，与“量子场是所有事物的核心”这一事实有着深刻的联系。

无即是有

认同这一事实的另一个结果就是，“什么也没有”实际上并不代表什么都不存在。换句话说，即使在一个完全没有物质(例如电子、原子)甚至光(光子)存在的空间里，仍然具有可测量的特性。这个空白的区域被称为“电磁量子真空”，是一种所有可提取的能量都被尽数去除的宇宙状态。与很多人的想象不同，它其实是一个容纳着很多活动的“大熔炉”，由波动的场组成却不包含任何光子。令人惊讶的是，量子真空中能够产生可以被观察到的现象。这不禁让我们发问，怎么能从“什么也没有”中产生出我们可以观测到的现象呢？

我们已经知道了，光可以被认为是电磁场的波动。请将这种波动所产生的场想象成海面上的涟漪。这些波动可以连续冲击海面上的任何船只，但是并不会将船上下移动或者将船推到某一条明确的波浪上。总的来说，船在这样的波动下并不会发生移动，仅仅是来回摇摆而已。现在将带电粒子(如电子)放在同样的想象场景中，这个带电粒

子可以“感知”到电磁真空中的随机变化，并受到这些变化的连续冲击，就像海面上的涟漪一样。如果电子在原子中被束缚，那么这种冲击就是电子在其可能占据的量子态之间的能量的转变。由于原子吸收光子的频率取决于电子的量子态间的能极差，因此，通过观察原子可能吸收的光的颜色变化，就可以知道电子在不同量子态间的能极差。这种变化微不足道，不到光的波长的十亿分之一。尽管非常微小，但是利用精确的频率测量技术，还是可以确定这种变化。20世纪50年代在纽约工作的威利斯·兰姆(Willis Lamb)是完成这一观测的第一人，并因此获得诺贝尔奖，他观测到的这种频率改变被称为“兰姆移位”。

对光双重身份的理解有许多层面。即使在前量子世界，光到底是波还是粒子的二元对立问题也需要解决。当时，通过了解光所体现的波动性质，以及与光相互作用的物体尺寸和性质，这一对立得到了解释。事实上，当物体的尺寸远大于光的波长且不具有锐利边缘时，物体的运动都可以解释成粒子沿着明确定义的路径做运动。而量子力学则提供了一种解释这种二元性的新观点。当光与物质相互作用时，光被视为一个或多或少带有能量的粒子，与此同时，它还保留了可以同时展现波动现象的能力。这个解决二元对立的方案引入了一个全新的概念：量子场。光粒子，即光子，是由量子场所激发，并且根据麦克斯韦量子版的光波方程进行传播。

量子场现在被认为是构成宇宙的基本实体，支撑着所有类型的物质和非物质，其中光可能是量子场最简单的一个例子。对此唯一的解释是，世界上的事物既不是粒子也不是波，而是兼而有之。这就是真实世界的本质。

[1] 作者误将杰弗里·泰勒的姓名写作George Taylor。

05 光物质 Light Matters

光是如何产生的呢？在回答这个问题之前，让我们先看看身边多种多样的发光物体：首先是日常生活中的灯具：基于金属丝发光的普通灯泡；荧光管；我们之前提过的激光笔；从烤面包机到汽车仪表盘等电子设备上的指示灯；阳光，当然还有星光；在地球南北极地才可以有幸看到的**极光**；不仅如此，还有萤火虫、萤科虫类以及船尾的磷光[1]等等。这些千差万别的物体是通过什么样的方式产生了一个共同的产物——光的呢？

这个问题的答案是，它们都涉及物质，更具体地说，都涉及电荷的转移。当这些电荷加速时，也就是说当它们改变运动速度或者方向时，就会产生光。这是一个简单的物理原理，对它的认识是电磁学理论的伟大成就之一。电场的起源是电荷，比如原子中的电子，其所产生的电场会延伸到整个空间，并吸引像质子一样的带异性电荷的粒子，且这种吸引力会随着其与电子间距离的增加而迅速减弱。正如我在第3章提到的，这是静电产生的力。

振荡原子与弯曲电子

现在假设电子突然运动起来，它周围的场也必定会随之改变，这是因为两者间有着千丝万缕的联系。图25描述了这种电场的变化，它看起来像一个“扭结”。位于电场中的质子并不会立刻感应到这种电场的变化，事实上，从电子发生运动到质子感知到电场变化，这之间有一个时间差。在这期间，“电子发生了运动”这一信息是以光速由电子向质子进行传播的。当质子感知到变化之后，质子会根据电子的运动方向而做出反应：如果电子靠近质子，那么质子受到的电场会变

强，从而受到更大的力；如果电子远离质子，则质子受到的电场变弱，从而受到的力减小。

现在假设电子来回运动，它周围的电场也会随着这种振荡同步发生变化，并且传播到质子所在位置，质子受到这种变化的电场作用发生振荡。振荡的电场(以及相关的磁场，这里我们不展开讨论)正是我们所说的光。

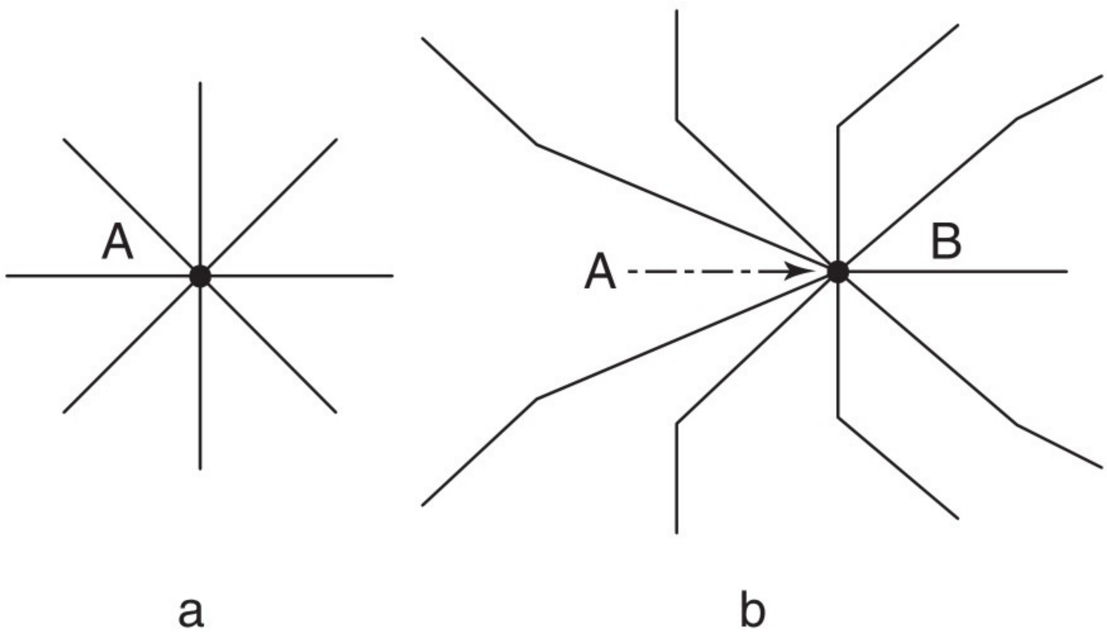


图25　a.静止电荷的电场线；b.加速运动电子的电场线。当电子加速运动时，如图中电场线的转折(或“扭结”)所示的电场的变化，会以光速远离电子

由于最简单的氢原子只包含一个电子和一个质子，所以我们可以由氢原子入手，理解原子是如何产生光的。首先，让我们考虑一下，当一束光照射在基态原子上时会发生什么。光迫使原子内部的带电粒子——电子和质子发生运动。但是由于电子比质子轻得多，在给定相同作用力的情况下，电子更容易运动，所以我们可以把质子看作近似

静止的，只考虑电子相对于质子的运动。事实上，电子以光的电场频率振荡，并随着电场的变化交替进行加速或减速。

这有点像推孩子荡秋千的过程。使秋千荡起来的最好方法就是按照秋千的自然振荡周期同步推进，也就是说，在秋千每次运动到最低位置时推一把孩子。即便如此，想让孩子在秋千上荡得足够高也需要费一番力气。孩子荡到最高位置时的加速度最大，而在最低位置处的速度最大。原子内部的电子也是如此，光束的能量被原子吸收，并转变为电子的运动。

假设你现在停止推秋千，会发生什么呢？孩子会以越来越小的幅度呈弧线摆动，最后慢慢停下来。同样，原子也是如此。电子逐渐停止振荡，并把自身的动能转换为光能辐射出去。这就是光的辐射过程，也是例如霓虹灯、荧光灯和激光笔等众多光源发光的原理。

在上面的说明中，我假设原子内部的电子之所以发生振荡，是源自一些光束的照射，才导致了光的辐射过程。从某种意义上来说，这就引出了一个问题：最初的这些光是如何产生的呢？事实上，我们可以用其他方法来“激发”原子从而产生光。例如，人们可以简单地通过加热材料来达到这一目的，普通灯泡就是这样一个例子。让电流通过金属丝，金属丝会被加热到很高的温度(有几千摄氏度)。随着金属丝的温度越来越高，电子开始与原子互相碰撞，且碰撞的次数越来越多。这种碰撞激发了原子，也使得电子迅速地加速或者减速。通过这一过程产生的光具有很宽的颜色范围，具体会产生什么颜色的光取决于材料被加热后所达到的温度，而不取决于构成这种材料的原子类型。

电也可以通过其他途径产生光。例如，在发光二极管(light-emitting diodes，缩写为LED，常用于显示器的制造)中，通过它的电流或者说是电子流，可以被原子直接捕获。这种产生光的方法比利用

热源产生光的效率要高得多。荧光管也是利用电流直接激发原子，但这发生在充满气体的灯管中。不仅如此，许多不同的化学或生物反应都可以释放能量，其中一些能量会以光的形式离开原子或分子。这还可以用来解释萤火虫是如何发光的。

正如我们之前提到的，加速度具有两个方面的含义：一个是速度大小的变化，正如之前提到的氢原子中电子和质子速度的变化；另一个则是方向的变化，哪怕速度保持恒定时也一样。加速度的方向变化在汽车转弯时很常见。当汽车转弯时，你被推到车门的一侧或者座位的一侧，并感受到有一种力使你随着车一起转向。转弯的速度越大，你感受到的力也越大。这表明你在加速，即使行驶速度并没有发生任何改变。

当带电粒子经历这种只改变运动方向而不改变速度的运动时，它们也会发出光。想象一下，一群电子被迫做圆周运动(你可以想象成它们被固定在旋转的车轮边缘处)时，由于这个角加速度的存在，它们会产生光。随着电子做圆周运动的速度不断增加，光的波长会变得越来越短，因而光子的能量也越来越大。以这种方式产生的光称为**同步辐射**，这也是产生X射线的一种常用手段。它还与南北极极光的产生有关：来自太阳的带电粒子进入大气层时受到地球磁场的作用进行螺旋运动从而产生极光。

量子光的生成过程

以上提到的这些基本机制是所有光源产生光的基础。但是当一些原子以群体的形式进行活动时，它们的行动方式会对最终辐射出的光的特性产生强烈影响。正如我在第1章中所提到的，普通灯泡发出的光与激光笔发出的光截然不同。为了理解这一点，我们需要深入研究原子的结构，因为原子的发光过程并不能完全类比成孩子荡千秋的过

程。由于原子和分子是量子力学实体，因此对于原子的发光过程，我们需要考虑一下它们的量子特性。

就我们目前讨论的问题而言，原子或者分子的量子特性仅仅意味着它们中的电子只能保有固定的能级。如果使用荡秋千模型，这就意味着荡秋千的最大摆动幅度不能为任意值，相反，幅度被限定为几个特定的值，也就是说它的值是量子化的。更具体地说，秋千摆动的能量来自离散的“小包裹”或量子单元。当你推动秋千的时候，你只能使秋千以一个或者多个量子单元进行摆动。在原子内部，这意味着当电子吸收或者发射出单个光子时，电子的能量只能以上述的不连续单位发生跃迁。与日常生活中的能量标准相比，电子发生跃迁需要的能量非常非常小。你把房间里的灯打开，它消耗能量的速率是每秒60焦耳，即60瓦。而灯泡中的原子发射的单个光子的能量大约为10^{-18}焦耳，因此，一个灯泡每秒发出超过10^{18}个光子。

只要用适当频率的光照射在原子上，原子就会进入激发态，如图26a所示(当然，我们还可以用其他的方式激发原子，例如给介质通电流等)。现在，根据量子理论中电子的跃迁运动可以推论出，当原子中的电子处于任一离散能级时，都是非常稳定的，因而不会主动发射光。电子就好比是放在橱柜中某一层架子上的球，从理论上来讲，它可以通过掉落到下一层架子上来降低能量，但这在实践中是不可能发生的，除非你稍微推一下它，不然球自己是不会滚落下来的。

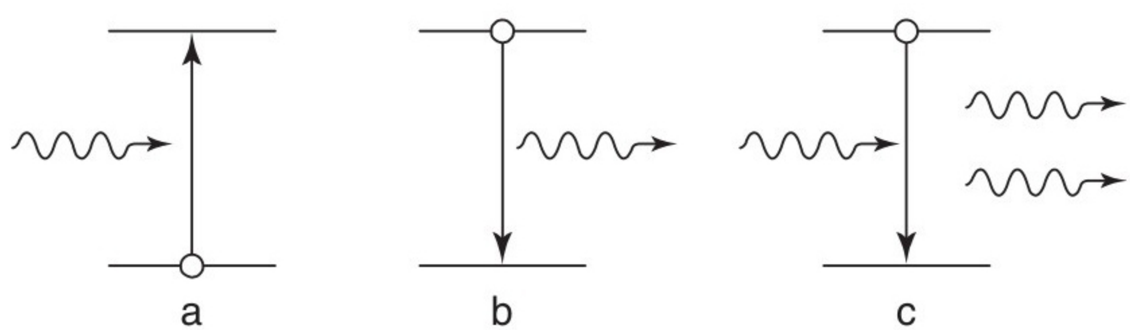


图26　一个原子吸收光子(a)、自发辐射(b)、受激辐射(c)的过程

这么看来，量子物理学似乎表明原子是不会发光的，因为你一旦把原子中的电子放在那些特定的轨道上，它们就会一直保持稳态，不会发光。事实上，除非电子处在最低能量状态，还存在一个推力使电子从较高能级跃迁到较低能级，否则原子是不会发光的。令人惊讶的是，这个推力是凭空产生的。

在第4章中，我提到了量子物理学最奇特的特征之一：即使是空旷的什么都不存在的空间，也充满热火朝天的“真空波动”。电磁场中的这些波动可以使原子中的电子跃迁到一个较低能级，并且，由于能级跃迁而产生的能量会以辐射光的形式释放出来。这个原子从较高能量的激发态跃迁到较低能量的基态，并释放出一个光子的过程，就叫自发辐射(见图26b)。每一个原子都能自发辐射，这一概念最初由爱因斯坦提出来，是为了解释光束与它所照射的物质之间的能量平衡关系。如果原子没有发生自发辐射，那么原子就会保持住来自光束的能量，我们平时到处都能看到的情况就不可能存在了，即大部分事物都处于一个与其周围环境平衡的稳定状态。

爱因斯坦认为自发辐射的核心奥秘就是它是一个随机过程。你不能确切地说出任何给定的原子何时会跃迁。你唯一可以说的是，平均来说，在一段时间(这取决于特定的原子，但是一般来说大约是一万亿分之一秒)内，在一个大的原子集合中，大约三分之二的原子会辐射光子。然而，这种基本随机性的起源一直是个谜，一直到1927年，保罗·狄拉克的量子场理论指出，量子真空波动是这种随机性产生的根源。他说一个根本不包含光子的场可以扰乱受激原子的稳定性，这一观点与我们的直觉相悖。一直到20世纪50年代，兰姆的测量才证明狄拉克的解释是正确的。

这意味着，即便是我们在日常生活中习以为常的现象，例如电视屏幕上的发光二极管生成图像，其核心也具有由量子力学产生的这种基本随机特性。与之相对，原子受到外界光的照射从而被迫释放能量辐射光子的过程被称为**受激辐射**。这种将原子的能量回收到光场中的方式并不具有随机性，这使得一种非常不同的光成为可能：激光。

相干性：步调一致

当原子和带电粒子各自“随意运动”时，如果数量较多，那么它们辐射出的光就会是一组不协调的波[2]。即使是尺寸小到只有1毫米的发光二极管，里面也有大量的原子，所以这种情况是很常见的。

这种发出不协调光波的辐射有一个特征，就是每个原子都会随机发射光子，与周围的原子在做什么无关。因此，光是向各个不同方向辐射的，且光子会在不同的时刻被发射出来。实际上，辐射过程的随机性表现在原子所产生光强度的随机性上。

假如将一个光子探测器放在灯泡前面，我们可以看到探测器输出的电流非常嘈杂，这是因为照射到探测器上的光的强度变化快速且随机，说明在每个时刻到达探测器的光子数也是随机的。

如何才能使原子步调一致，从而使光波具有相干性呢？我们可以回想一下之前的类比：想象有一些秋千，且每一个秋千的摆动频率都是相同的。这些秋千开始进行随机摆动，那么就会出现以下两种情况：第一种情况是这些秋千的摆动是不同步的，即在任何时刻，各个秋千都会运动到其固定轨迹上的不同位置；第二种情况则是这些秋千的摆动是同步的，即相邻秋千间的轨迹差异是固定的，就好比足球比赛中场上观众按照顺序依次站立和坐下形成的人浪。在第一种情况下，从这些不相干的原子发出的光就像灯泡或者发光二极管发出的光，这种光是不相干的。在第二种情况下，原子们产生一致的振荡，

它们产生的光以一种**相干**的形式发出，所有的光子都往同一个方向辐射。这就是受激辐射过程中会产生的情况(见图26c)，它是激光器的基础。

激光

激光器也许是20世纪光学领域最重要的发明。这个设备产生了非常有用的光束，革命性地改变了光的应用范围和性能。激光器可以作为一个特定的照明光源使用——例如在显微镜和光谱学中；不仅如此，激光器还提供了一种能将大量能量沿特定路径引导到特定目标上的手段，从而控制物质的动力学。这类应用的一个极端例子就是用激光驱动原子聚变，使得新形式的核能得以利用，从而提供大量的能量。我们将在第7章对其进行讨论。

激光器包括一个光学放大器，也被称为增益介质，其原子通过受激辐射产生光。它被放置在两个反射镜(可能还有其他光学元件)所组成的光学腔中。随着原子不断地发射光子，光学腔内的光子数会逐渐增加，直到原子释放的能量与由反射镜从光学腔泄漏出来的能量达到平衡。说得更详细点，当放大器被接通时，从放大器辐射出去的光会被光学腔末端的反射镜反射回来，这就进一步刺激了受激原子的辐射，从而使光学腔内的亮度增加。在另一端的反射镜处，一部分光作为有效输出从腔中透射出去，一部分光则返回增益介质继续激发原子产生光子。当光通过放大器进入光学腔的速率等于光通过镜子透射出去的速率时，我们就说激光器处于阈值。超过这个阈值，放大器增益的任何增加(原子进入受激状态的速率)都会导致腔内强度的增加，从而导致输出光的增加。

光学腔对于激光的颜色有一定的限制。结果表明，获得最大增益的频率是那些在每次往返过程中都会发生相长干涉的光波所对应的频

率。这意味着在腔内往返一次的长度应该等于半波长的倍数。满足这种谐振条件的频率即符合谐振腔模式。

激光器之所以重要，是因为它们发出的光是相干的：所有的光子都以相同的颜色朝着大致相同的方向运动。其中运动方向由光学腔决定，而颜色则由增益介质中的原子和光学腔共同决定。这就使得激光是以激光束的形式存在的，它与你想象中的“光线”非常相近。它在传播过程中仍然会因为衍射而发散，但是发散程度已经微乎其微。这种特性也意味着激光可以用透镜或者平面镜聚焦成一个特别小的光斑。

除了相干性之外，与灯泡发出的光相比，激光的第二个优势是它的颜色更纯。换句话说，激光发出的光的波长范围很窄，而灯泡发出的光的波长范围特别广。激光的光强度特别稳定(光探测器的探测结果中噪声很低)，可以连续发射或者作为光脉冲进行发射。

激光可以聚焦成非常小的光斑，这一能力使得它在显微镜技术中得到了各式各样的应用。例如，通过扫描在显微透镜焦点处目标物体上的激光光斑，并且检测从物体散射或者重新辐射出的光，可以构造出物体的三维图像。这个方法对于观察动物组织非常有用，这类光学显微镜在生物医学方面有着广泛的应用。

激光在制造业中的许多应用也是利用了这一特性。例如标记、切割、钻孔或者焊接金属等操作，都要求短时间内在金属的较小区域上聚集能量。高功率激光器可以产生以脉冲形式呈现的相干光束，聚焦能力强，因此成为这些材料加工操作的理想选择。

由于这些特性，激光在医学方面也得到了很多应用，涉及皮肤、牙齿和头发等材料，常见的如激光矫正视力和激光牙科技术等。激光也可以去除文身，利用激光加热皮肤里的文身墨水直到其被完全破

坏。激光也可以除毛，但很可惜的是激光并不能反过来让毛发再生。其他一些人们所熟悉的设备，如CD、DVD、蓝光光碟(Blu-ray Disc)和一些计算机磁盘存储设备，也都是利用了激光的聚焦性能卓越这一特性，使得材料能够储存密度非常高的数据。

激光的颜色可以非常纯，因而可以利用光谱法来区分不同混合物中的原子和分子种类。正如第1章所提出的，不同的原子或者不同的分子，由于它们的结构不同，因此具有不同的吸收和辐射光的特征频率。延伸一下我们本章中提出的类比，将这些原子和分子比作秋千，那么这些秋千中连接座椅的绳子长度不一，因此，这些原子或者分子的固有振荡频率由它们在混合物中的组合方式所决定。

事实上，每个原子和分子都有不同的辐射和吸收的频率范围，对应着不同的电子构型的激发。这些频率范围通常位于可见光光谱的蓝光区域，但有一些分子可吸收的光的波长要短得多，是人类不可见的。许多分子也可以吸收比可见光的红光波长还长的光。这种区别来自构成分子的原子核之间的振动。由于原子核比电子重得多，所以它们倾向以更低的频率振荡。这组频率数据相当于分子的一种“指纹”，可用来确认分子类型。

当然，用这些“指纹”所确立的目录在化学中是很重要的，因为通过它可以识别出化学反应所涉及的不同元素。“指纹”也被用于分子生物学；甚至在细胞生物学中，可以通过它来观察特定的标定分子，从而研究分子的运动。这种“指纹”对于天体物理学也相当重要，它可以确定恒星、星系、星云中存在的元素。在大气物理学和气象学中，则可以遥测污染物和颗粒。这种监测为评估气候变化的影响和成因方面提供了关键数据。

通过组合几个不同激光器所发出的激光，例如一个发出红光，一个发出绿光，一个发出蓝光，那么就有可能制造出一台激光投影仪。

根据电脑或者互联网输出的视频信号来改变各个激光信号的强度——也许可以通过液晶单元的方式实现——那么色彩鲜艳、高饱和度的电影就可以投射到屏幕上。红、绿、蓝这三色光的组合足以构成一个完整的调色板，而激光可以在屏幕上产生非常明亮的图像。**X射线**

当光的波长非常短，属于光谱的X射线区域时，就会产生一种特定的光谱。X射线的光子能量很高，不仅可以激发最外层的电子，也能激发原子中被束缚得最紧密的电子。这意味着X射线可以用来观察原子和分子的核心，并了解它们所处的环境，由此，我们就可以改变电子结合能[3]。X射线的吸收光谱技术在材料研究中有着广泛的应用，从检测微量污染物到研究玻璃的结构。如第3章所述，X射线衍射法也被用来研究晶体的结构。当X射线的波长接近于晶体中原子的间距时，晶体就充当了“衍射光栅”，使X射线向离散的方向散射。通过监测相机上的衍射图样，利用先进的反演算法可以重建高度复杂的晶体的三维结构。从分离出的生物和化学分子中提取特征，确定可能的新分子结构，对比进行设计，从而实现某一特定功能，这在现在已经是一个常规的流程。

同步加速器可以为这种光谱技术提供最好的光源。为了产生符合要求的具有短波长的X射线，同步加速器必须产生高能电子束，并且在一个环形轨道上对它进行加速。随着电子的加速运动，实验站会捕捉到闪烁的光，这导致了X射线的短暂暴发，可以用于衍射成像。例如英国哈维尔的“钻石光源”[4](Diamond Light Source)，电子束在一个超过500米长的环内被加速至超过十亿伏特。下一代X射线光源正在使用线性粒子加速器来制造，这种加速器可以产生极其明亮的X射线光束。如图20所示的X射线衍射图就是用“钻石光源”拍摄的。

超短光脉冲

激光也能够以短脉冲的形式发出，方法不止一种。能够产生最短光脉冲的方法称为**锁模法**。锁模法要求增益介质具有较大的带宽，也就是说，它可以在比较宽的光谱范围内对光进行放大。这使得光学腔内多个不同种类的激光都可以获得增益。如果让这些不同种类的激光具有相同的相位，那么拥有不同频率的光波将在光学腔内叠加形成单个脉冲，并且在两面反射镜之间来回反射。脉冲有多短取决于锁定频率的数目——频率分布的范围越宽，脉冲越短。

超短激光脉冲的获得使得时间分辨光谱技术成为可能。这种技术让我们得以利用闪频仪看到事物是如何随着时间变化的。利用光来“定格”快速运动的这一想法，可以追溯到19世纪末埃德沃德·迈布里奇(Eadweard Muybridge)的工作。他萌生了利用快速照相机快门来拍摄骏马奔跑的想法。由于马腿移动得太快，人眼无法进行分辨，以至于当时人们还无法分辨马在跨步的过程中四条腿是否在某一时刻同时离地。为了解决这个问题，迈布里奇沿着跑道设置了许多相机，每一个相机快门都由一根横跨于跑道的线所操控，当马奔跑经过这些线时就会依次触发快门。这使得他能从马的反射光中提取出一个很短的光脉冲，这种光脉冲持续的时间比马腿移动的时间还要短。他这项研究的目的与成果就是告诉他的研究资助者利兰·斯坦福(Leland Stanford)，在马奔跑的过程中，马的四条腿在某一个瞬间可以同时离开地面。

传统相机的机械快门虽然可以很快速地关闭，但是仍然不够快，因此依然无法观察到某些形式的动物运动，例如蜂鸟翅膀的扇动。至于一些更快速的物理过程，例如发生在千分之一秒的时间尺度上的爆炸，这么短时间内发生的变化是无法用传统相机捕捉到的。为了解决这个问题，麻省理工学院的哈罗德·埃哲顿(Harold Edgerton)在20世纪50年代发明了一种基于光学开关的新型非机械式快门。他可以用这种装置拍摄爆炸事件的静态照片。

这些开关我们可以称之为“被动”仪器，它们在打开状态时有一片允许光通过的空间，所以这种设备适用于照明良好的物体(例如沐浴在加利福尼亚阳光下的马)或者是本身就发出大量光的物体(例如爆炸)。我们可以想象一种“主动”的仪器，它可以产生短的光脉冲来照亮一个移动的物体，例如照相机的闪光灯所发出的光脉冲。与物体移动所需要的时间相比，一道持续时间更短的光照射在物体上时，就可以观察到某个瞬间物体定格住的图像，即使快门的速度比物体的运动速度要慢。第二个闪光会定格稍后的运动图像，依次往后推，这样就可以获得物体运动的动图了。

将在同一事件的重复实验中拍摄的序列帧组成一部电影可以说明运动对象的快速变化，即使其变化的速度已经远远超过人眼的识别程度。确实，这些稍纵即逝的事件竟然能够通过这种方式被观察到，真的很令人惊叹。埃哲顿于1931年发明了“频闪仪”。他的一些最具代表性的图像，例如子弹穿过苹果或者扑克牌(见图27)，都是用它拍摄的。



图27　频闪仪记录下一张图片，显示了运动中的子弹被定格的瞬间

使用现代脉冲激光作为“闪光”，不仅可以观察到运动的子弹定格的瞬间，还可以观察到参与化学反应的分子中原子的运动[5]，甚至能看到在原子核周围以更快速度运动的电子。这些运动的时间尺度小得惊人。对于分子来说，它的时间尺度小于1秒的一万亿分之一（100×10^{-15}秒，即100飞秒）。而对于原子内的电子来说，其时间尺度为100×10^{-18}秒（100阿秒）。飞秒化学[6]和阿秒科学[7]都是研究光和物质相互作用的前沿领域，我在第7章将会作进一步讨论。

[1] 当波浪或者船只扰动了生活在海水中的某些微小浮游生物时，它们会发光。因此，这种磷光其实是一种生物发光。

[2] 例如波的方向、相位不一致等。

[3] 电子由被束缚状态转移到无穷远时所做的功，可以用来衡量电子被原子核吸引的紧密程度。

[4] 位于英国牛津郡，是英国第一台第三代同步辐射光源。

[5] 因为这一发现，1999年的诺贝尔化学奖颁发给了艾哈迈德·泽维尔(Ahmed Zewail)。

[6] 研究飞秒时间尺度内的化学反应过程和机理的一个领域。

[7] 阿秒时间尺度上的科学研究。

06 光、空间与时间 Light, Space, and Time

对于罗伯特·格罗斯泰斯特(Robert Grosseteste)来说，理解光的本质是理解整个世界的钥匙。格罗斯泰斯特曾担任英国林肯郡主教，也是英国牛津大学在13世纪迎接的第一任校长，他非常崇尚古希腊人的作品。对他来说，就像许多哲学家一样，理解光的本质是理解世界的关键。他在论文《 论光》(De Luce)中是这样颂扬光的重要性的：“在我看来，光……是最首要的物质形态。光可以沿任意方向传播，因此一个发光点向各个方向传播的光可以瞬间形成一个任意尺寸的‘光球’。”

对于格罗斯泰斯特而言，光可以用来定义空间：光在传播途中所经过的区域即为空间。在他看来，如果没有光，就不会有空间，因此，物质以及物质所在空间的定义不可能脱离光的概念而存在。基于这个信念，格罗斯泰斯特还将光、空间与物质这三者之间的密切联系作了量化。这样的观点在随后的几个世纪里影响了宇宙学的发展。

时空

牛顿认为，空间这个概念既不需要被承认，也不需要被定义。他认为空间是一个早已存在的实体，就像是一个舞台，只等着演员上台表演精彩的剧目。因此，他提出的一系列力学定律中，很多都与物体在空间中的大规模运动密不可分。与牛顿不同，爱因斯坦则将光放在理解空间概念的核心地位。他认为可以通过信号从宇宙中一处发送到另一处的速度限制来定义时间和空间，正因为光速是一个确定的值，所以空间和时间这两个概念不可分割。爱因斯坦的相对论告诉我们，不能单独考虑空间或者时间中的一个，因为我们认知空间与时间时，都是利用对其进行局部测量的方式，但如果在一个正相对我们运动的

系统上测量同一空间与时间，测量结果将发生改变，这也是源自光速的确定性。

光与时间、空间这种奇妙的关联是怎么产生的呢？让我们从牛顿空间的概念开始说起。让我们想象一个脚手架，如图28所示，固定长度的短棍相互连接形成了一个三维立体框架。牛顿认为，这样的结构早已存在于所有事件发生之前，事实上，所有事件都发生在该结构的某个地方。因此，只要在结构中确定一个参照点，事件的位置就可以通过参照点与事件之间的短棍数来确定。那么怎么确定事件的时间呢？在脚手架每个短棍的衔接处放置一个时钟，保证这些时钟在脚手架的任何位置都显示相同的时间，这样我们就可以很容易地确定事件的时间了。这样的时间被称为“通用时间”。

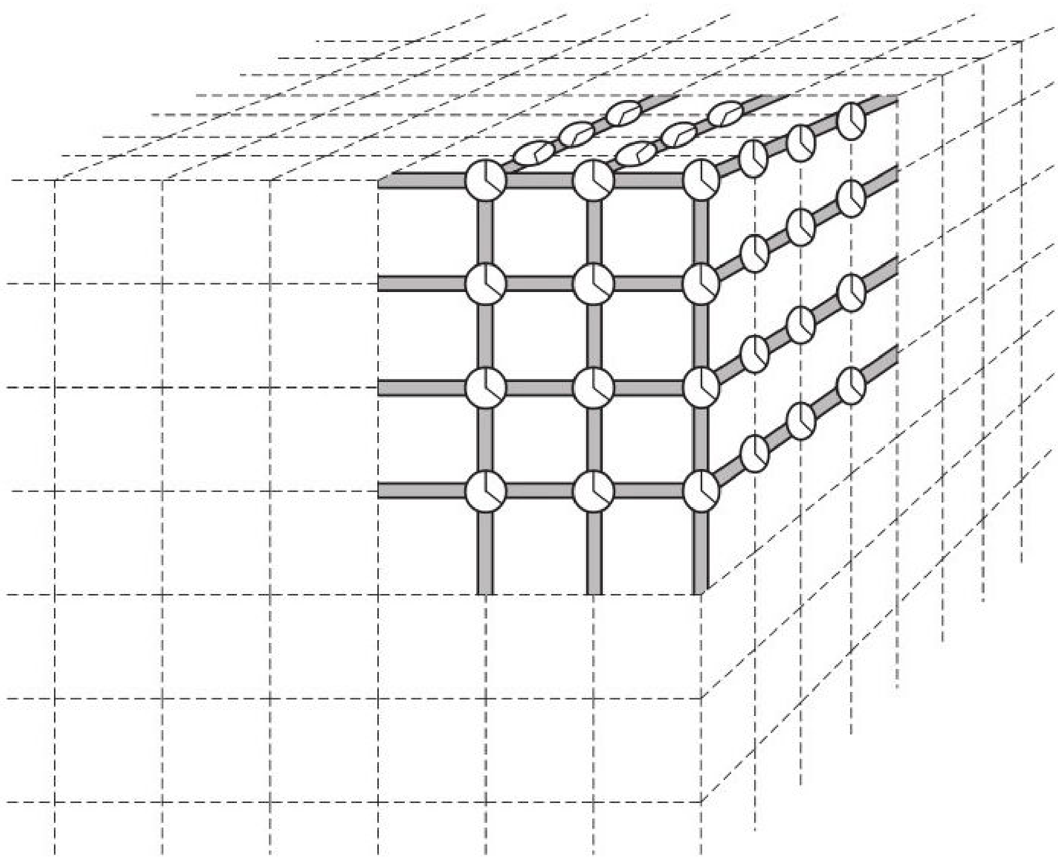


图28　图中的“脚手架”展示了一个代表空间的三维模型，每个格子的边由一个可测量距离的短棍表示。每一个节点上都有一个时钟，且都是同步的

现在，我们必须提出这样几个问题：首先，我们应该如何造一个可以计时的时钟？其次，我们应该如何确保它们在空间中保持同步？最后，我们应该如何确立长度标准呢？这些问题的答案都与光的属性密切相关。实际上，我们可以用光来定义长度：一米是光在1/299 792 458秒内在真空中行进的距离，这就回答了上述的第三个问题。由于长度的定义与时间息息相关，那么就回到了第一个问题：我们造出的时钟到底可以有多精确？

时钟

时钟最重要的特征是它以规则的时间间隔发出信号，并以此为周期不断重复下去。通过计数两个事件之间相隔的信号数或周期数，我们可以计算出事件之间的时间间隔。时钟越精确，就意味着每两个相邻信号之间的时间间隔越规则，周期越精准。座钟是一种机械钟，它的周期由钟摆的左右摇摆间隔确定，但是其精度很容易受到温度和湿度的影响。相比而言，电子手表的周期依靠的是石英晶体的振荡，其振荡频率比钟摆的摆动周期更加规则，因此石英表计时一般比座钟更加精确。

世界上最精确的时钟是原子钟，它的计时功能与原子中电子的运动息息相关。我们知道，围绕在原子核周围的电子排列在不同的轨道上，因而存在不同的能量层级。通过吸收或释放某些特定频率的电磁波，电子可以在能量层级之间发生跃迁，即从一个稳定的能量层级跳跃到另一个稳定的能量层级。对于某些原子，比如铯原子，它发生电子跃迁时两个能量层级之间的能量差已经被精确定义了，因此只需要观测能量差出现的频率，即使得电子发生跃迁的电磁波频率，据此就可以定义时钟的周期了。

在实际操作中，我们使用微波照射铯原子，再慢慢改变微波的频率，直到电子开始在两个稳定的能量层级之间进行跃迁，我们就可以通过这一频率来定义时钟的周期。

构建这样的原子钟面临着很多技术挑战。首先，我们需要将原子冷却，使它们回归理想的初始状态；然后，我们要谨慎选择微波照射的方式，以最大化原子钟的灵敏度；最后，我们还要能够精准确定电子是否已经跃迁到较高能量层级的状态。现在基于铯原子的时钟已经成为测量时间的最准确方法，它在一亿年内的误差小于百万分之一秒。

原子钟提供了一个受国际肯定的时间标准，由一些政府实验室维护，例如美国国家标准与技术研究院(National Institute of Standards and Technology，缩写为NIST)、英国国家物理实验室(National Physical Laboratory，缩写为NPL)和德国联邦物理技术研究院(Physikalisch-Technische Bundesanstalt，缩写为PTB)等。时钟是很多技术的基石，例如，它们对于制定全球定位系统(global positioning system，缩写为GPS)的统一基准至关重要。现在，GPS已经运用到日常生活的方方面面，例如汽车的卫星导航系统等。可以说，时钟在我们的生活中几乎无处不在。

时钟的同步

下一个挑战是如何使两个时钟同步，使得它们可以被统一校正。其中一种方法是将信号从一个时钟发射到另一个，通过测量信号的延时来完成校正。具体做法是首先启动第一个时钟，当该时钟完成第一个时钟周期时，向另一个时钟发射一束光。管理第二个时钟的人收到光信号时，可以通过对比知道第二个时钟与第一个时钟间的延时情况，并利用这些信息进行时钟校正。由于时钟的构造都是一样的，我们假设它们计时周期也相同。

这样做会有一个有趣的结果。设想一下，为了让一个地球上的时钟与另一个放在遥远星系中某一颗行星上的时钟同步，你把一束光发射向那个行星，即使光速很快，由于距离实在太遥远，光到达那里还是要用很长时间。与此同时，你却在慢慢变老。所以，行星上的人接收到的这一校正信号，其实是你年轻时发出的信号，他看到的你也是很多年前发出信号时的你。

同理，当我们仰望星空，我们看到的实际上是源自很久很久以前从遥远恒星表面发出的光。当我们遥望更远的恒星和星系时，我们看见的是更久远的过去：一个数十亿年前的宇宙。从这个意义上说，我们接收到的光也有数十亿岁了，从它诞生的那一刻起，就一直在宇宙中穿行。光是我们在宇宙中能看见的最古老的东西。

当然，我们通常所说的时钟相距没这么远。有这样一个有趣的现象，如果你把一个时钟放在一架飞机上，以大约每小时800千米的速度飞行，你会发现，它相较于地面上的时钟要走得慢一些，这是因为信号在两个时钟之间得以传递的最大速度是光速。

看图29你就能明白其中的道理。一个人(标记为A)在地面上，而另一个人(标记为B)在高速运动的飞机上。A向离地面H高度处的一面镜子发射一道光。从A的角度看，这道光的行程是H；然而，从B的角度看，这个光信号的行程比H稍长，因为A相对于B在高速后退。既然信号的传播速度对A和B而言都是光速，而且他们根据发射信号和接收信号所记录的时间是一样的，那么唯一的解释就是从A的角度看，B的时钟走得比自己的慢，而从B的角度看，A的时钟走得比自己的慢。这个现象被称为**时间膨胀**。

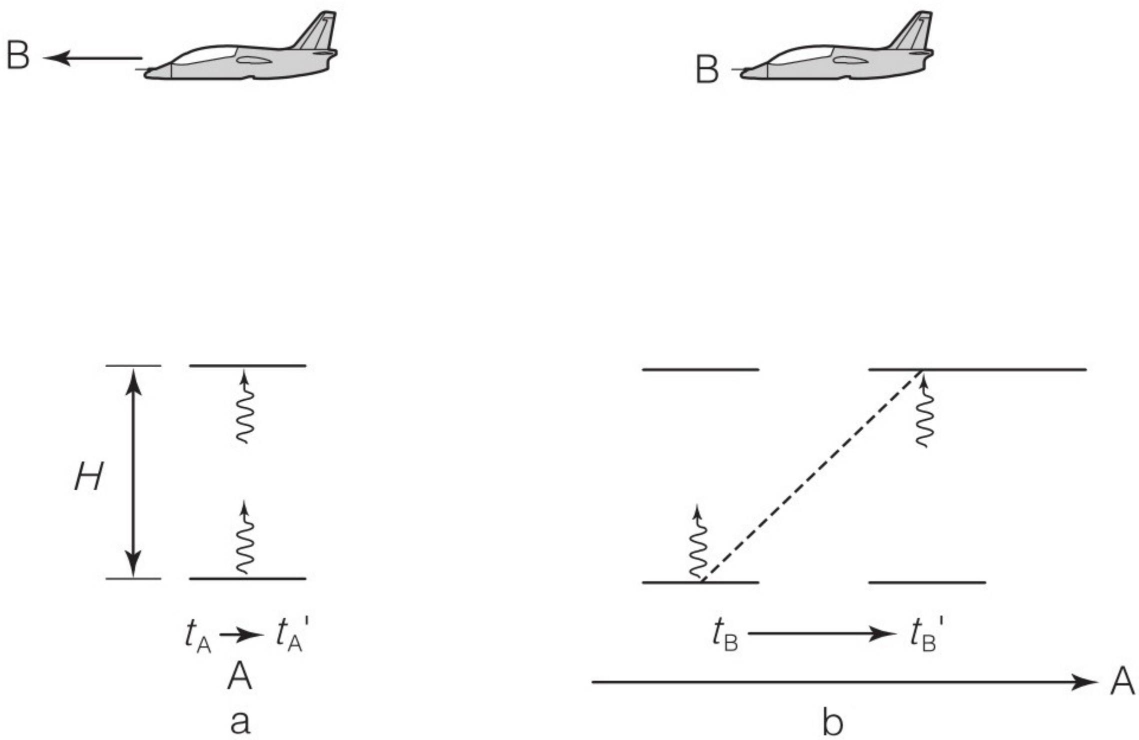


图29　相对运动导致的时间膨胀。观察者A(地面上)和B(高速运动的飞机上)各自测量同一束光到达悬挂于高H的镜子的时间。他们测出的时间并不相同，这是因为他们之间有相对运动

爱因斯坦用同样的方法表明了空间也同时在收缩。也就是说，对观察者A而言，观察者B所处三维空间框架中的短棍(见图28)应该比A的要短，反之，B也觉得A的比自己的短。

这两个效应的产生是因为任何信号都有一个传播的最大速度，而且这个速度对每个人都是一样的。如果不是这样的话，那么我们就可以自行选择一个三维空间框架，或者说“参照系”，在这个框架中，信号以最快的速度传播。爱因斯坦的相对论指出了这个框架是不存在的，因此，牛顿提出的空间是恒定不变且预先存在的想法是不正确的。由于任何信号传播的最大速度都是光速，所以光对于定义空间和时间都至关重要。

那么，使用之前提到的“光脉冲”方法去校正两个时钟到底能达到怎样的精度呢？一个提高校正精度的方法是让这束光尽可能地短，这样信号到达时的不确定性会最小。因此知道光束的长短是否有极限很重要。事实证明，这些极限确实存在，并且它们源自一种类似波的特性，这种特性限制了成像系统的分辨率，正如我们在第3章讨论的那样。

那么，光波的频率是如何确定的呢？想象这里有一个时钟，观察在单位时间间隔内我们可以接收到多少个波峰。通过的波峰数量越多，表明光波的频率越高。由于我们判断光波是否达到波峰的能力并不完美，所以频率的测量精度取决于我们重复测量的次数。如此说来，我们观测的时间越长，对频率的测定就越精确。值得注意的是，在此过程中，时间的测量也存在着不确定性。这种时间与频率测量精度间的权衡是波的基础：频率的不精确性与时间间隔的不确定性相乘为定值。19世纪的法国数学家和科学家约瑟夫·傅里叶首先指出了这一点，他在建立光传播的波动模型方面起了关键作用。

超短光脉冲

傅里叶的定理对于时钟同步很重要。这一定理指出，如果要产生一个历时很短的短脉冲，我们必须有一个不确定的频率。换一种说法就是，一个短脉冲是由多种颜色的光构成的。这完全类似于阿贝所定义的光学成像的情况：为实现高分辨率成像，我们需要小焦距，需要收集来自较大范围角度的光线。正如阿贝所揭示的，光学成像分辨率的最小尺寸可被视作约等于光的波长，顺着这个类比推理，傅里叶表明，光脉冲最短持续时间应该就是光波的一个周期。

这在实践中意味着，在电磁频谱中的可见光区域，可以产生持续时间约为2飞秒的脉冲。令人惊讶的是，从激光中获得的光源现在已经

可以成规模地产生这样的超短光脉冲了。它的实现是基于第5章中所描述的激光锁模技术。

尽管持续时间已经非常短了，它却并不是自然界中存在的最短光脉冲，甚至也不是实验室中可以产生的最短光脉冲。事实上，平均波长更短的光源产生的光脉冲会更短，这是因为若将波长缩短，光学周期也会跟着缩短，原则上就可以缩短脉冲的持续时间。用这种方法产生的光脉冲在目前保持着最短可控光脉冲的世界纪录。在具体操作上，我们将非常强的激光照射到原子气体上，经过被称为“高次谐波产生”的过程，一个频率是激光频率几十倍的光波就这样产生了。这样的光脉冲的持续时间为几十阿秒。这个数字让人难以置信，因为这一持续时间等于电子在原子内振荡所需的时间。

频率梳

第5章中我已经指出，在锁模激光器中，单个脉冲需要穿过一个光学腔。每当它遇到一个平面镜并发生反射时，其中一小部分光脉冲就会穿过平面镜并从腔体中射出。因此，在光学腔外部，光看起来就像一个“连续的”脉冲序列，且这些脉冲之间的间隔等于光在腔内的往返时间，通常是十亿分之一秒左右。观察这样的脉冲序列(见图30)，你会发现相邻两个脉冲之间的时间间隔要比这些脉冲的持续时间长得多，看起来就像梳子的齿一样。通过仔细调节产生脉冲的激光，可以使各个脉冲彼此相同，从而保证每个脉冲的电场在同一时间达到峰值。

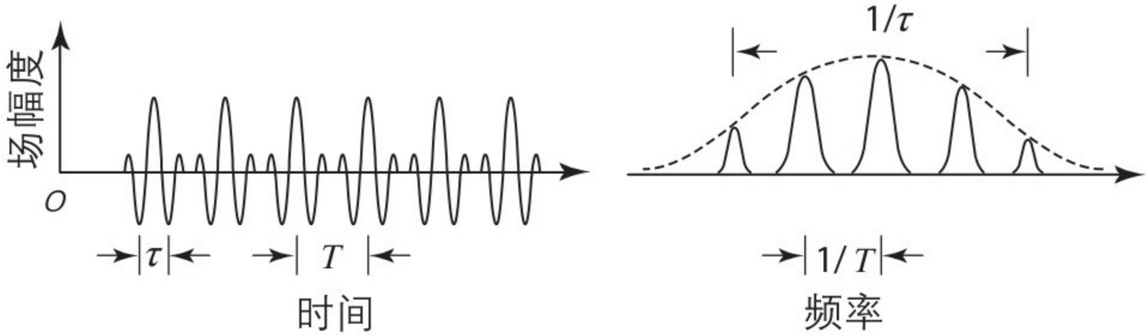


图30　一系列几乎一模一样的单周期光脉冲。脉冲序列的频谱看起来像梳齿，因此被称为“频率梳”

频率梳中的每个“齿”在绝对频率处都有一个非常精确的位置。一组精确校准的频率对于构建精确时钟非常重要，因为它使得我们可以直接将光学频率与其他较低频率(通常是微波)比较，这种比较一般通过电子设备来实现。

使用这一方法，我们可以将铯原子钟的频率(处于微波波段)，与像锶原子或铝离子由于电子跃迁而产生的处于光学波段的频率进行直接比较。例如，现在卫星导航中使用的标准铯原子钟，可以实现时间同步并以相同的时间间隔进行计时，其一亿年内的误差小于百万分之一秒(误差在1×10^{-18})。这一结果是由锶或铝中的光电子振荡频率的精度所决定的。

这种“光学钟表”可以非常精确地比较不同的频率，因此可以提供一种测试相对论原理的方法，从而使我们更好地理解光在定义空间和时间方面的作用。到目前为止，在所有物理量中——频率，也就是时间，是测量精度最高的。

光通信

频率梳在基于光的电信链路中也很重要。在第3章中，我们已经知道光波可以沿着光纤或玻璃“芯片”传播。长途电信基础设施就是利用这一现象，连接起世界各个角落的居民，同时它也是互联网技术的基础。光的通信之所以能得到如此广泛的应用，是因为与电线甚至微波蜂窝网络相比，它承载信息的能力更强。这使得大规模数据传输成为可能，比如需要通过因特网传输视频的情况。

许多电信公司提供“光纤宽带”服务。这些服务把宽带的速度作为关键卖点，宣传可以加速到每秒100兆字节(Mbps)来接收和传输数据。一个字节(byte)包含8个比特(bit)，而每一个比特只能是数字1或者0。信息以“0”和“1”组成的序列通过光纤发送，最终由计算机或移动电话将这些“0”和“1”的数字串解码为可以被人们轻松理解的视频、音频或文本消息。在光通信中，每一个比特的“0”或“1”的信息由光束的强弱表示：通常低强度为0，高强度为1。每秒送达的这些“0”和“1”的信号越多，通信速率也就越快。电信公司服务中说的Mbps，就是我们在该公司链路上传输和接收信息的速度。

为什么光通信这么好用呢？主要有两个原因。首先，光束不容易相互影响，因此只需要一根光纤就可以同时传输许多种光脉冲(通常颜色不同)，而不用担心信息发生混杂。这是因为制造光纤的玻璃不吸收光或仅吸收极少量光，因此不会加热和破坏其他脉冲序列。

此外，只有当在玻璃中传播的光强度非常高时，才能影响另一束光。举个例子，当两个激光器分别发出光束，即使这两个光束相互交叉，也不会出现光束扭曲或偏离其原始路径的现象(即使激光器具有巨大的功率，在真空中观察时也不会出现这样的现象。但是如果你在非真空的环境下观察，由于光的传播路径中充满空气，光束会出现扭曲或偏离其原始路径的现象)。这意味着在大多数材料中光束间的“串扰”非常微弱，因此可以同时传输许多光束而不会导致信号变差。这

与沿着铜线传输电子的情况非常不同，它是在本地搭建“有线”通信链路的常用方式。由于电子通常会加热电线，导致电子自身的能量发生耗散，使信号更难被接收。因此，在这种情况下，不同信号的通道数量必须保持足够小，才能够避免这一问题。

采用光学通信的第二个原因是，光波会以非常高的频率振荡，这使得超短脉冲的产生成为可能，这一点我们在之前已经讨论过了。这意味着脉冲与脉冲之间的间隔可以非常小，因而每秒可以传输更多比特的信息。实际上，现在的商用长途系统已经可以达到40 Gbps的速率。相比之下，铜线中的电信号由于之前提到的热效应，使得带有信息的脉冲在持续时间和时间间隔上都受到限制，并且这种热效应会随着频率的增高而加剧，为信息的传递带来更多的限制。可以说，铜线在以比光纤通信传递速率低得多的情况下就已经筋疲力尽了。

基于光纤的光学通信网络可以支持色域非常广泛的光，这是因为能被玻璃传送的光的波长范围较宽。尤其是1.3~1.55μ m波长范围内的光在光纤中几乎不发生散射或吸收损耗。在这些波长下，光子在光纤中的耗损率约为每千米5%。这样微小的损耗可以通过在光纤中放大光的强度来弥补，从而使超远距离传输(例如横跨大西洋)成为可能，无须将光转换为电信号，或是将电信号转换为光。

电信频谱窗口被分成许多单独的频率“槽”，类似于图30中所示的频率梳。每个频谱分量是一个单独的通信信道。频率窗口中可以有150个左右这样的槽。在每个通信信道内，可以操作40 Gbps的光信号。因此，通信的总比特率为150乘以40 Gbps，即6 Tbps。[1]

随着互联网的普及及其层出不穷的各种服务，我们对通信带宽的需求也在不断增长。这促使光学工程师们不断开拓创新，探索如何更充分地利用光的潜力。

[1] 1 Tb， 1太比特(Terabyte)，等于1万亿比特。

07 探索光学前沿 Lighting the Frontiers

光学有着悠久的历史，尽管它可能是自然哲学和科学中一直持续发展的最古老分支，但是时至今日，它仍然处在研究和应用的前沿。光学无处不在：它不仅可以作为检测、成像和通信的工具，也提供了探索、发现和解释新的基本效应的方法。

利用光可以创造出一些物理学上的极端条件，例如自然界中不存在的极端温度、极端压力和应力。这些极端条件或许也存在于最遥远的恒星中。光还可以用来观察甚至控制发生得极快的事件，例如发生在原子内部的电子运动。

此外，光还可以展示量子世界的奇特特征。它揭示出，即使在日常条件下，这个非连续的世界中也存在着许多与直觉相违背的方面。而这个非连续的世界却构成了我们日常经验中稳定可靠世界的基础。在这一章中，我将探索一些由光引领的前沿领域。之所以能够研究这些领域，是由于在光源、光学系统和探测器方面取得的巨大科技进步，使我们能够在空间和时间上精确控制光束的形状和强度。

光力学

光能够对物体施加作用力，这使得我们可以利用成形光束对小块材料进行“远程控制”。光可以用来移动物质，使其与其他物体接触，或者用来操控分子和原子的内部结构，迫使它们发生简单的化学反应，从而可以研究和开发具有特殊性质的材料。光的这种功能在很多研究领域都非常有用。

光之所以能够产生机械力，是因为光的每个光子都携带动量。例如，当光子从平面镜反射回来时，平面镜会受到一种力，这种力帮助

光子改变了运动方向。这就像是消防水带中的水撞击墙面，在反弹回来时向墙面施加力一样。

类似地，当光子发生折射时，它的运动方向会发生改变，这也需要力的帮助。所以光子会对折射元件施加力。如果一束光入射到玻璃珠上，其中与玻璃珠几乎相切的光线的方向改变最大。当光线穿透玻璃珠表面时，穿过玻璃珠下半部分的光子会向上运动。因此玻璃珠会受到一个反方向的力。随着光子的运动方向改变，光子在前进方向上(遇到玻璃珠之前的运动方向)的动量减小了，这说明光子对玻璃珠在前进方向上也有一个净力。这个力的强度取决于玻璃珠每秒折射的光子数。最终，如果这个光束中心的强度强于光束外围，那么玻璃珠就会被推向光束强度较高的部分，就好像光束将玻璃珠“捕获”了一样。

这个效应可以将聚焦的光束变成一个“光学镊子”。光学镊子能够抓住微小的物体，并且通过操控光束方向来控制物体的移动。光学镊子可应用于生物学，例如，它可以操控单个DNA链的位置和运动，还可以用来研究分子马达[1]的特征。具体来说，DNA、蛋白质和其他重要的生物分子都可以附着在这些玻璃珠的表面，因而光学镊子可以依照上述原理对它们进行操控。光学镊子可以在比光的波长还要小的精度上控制这些分子的位置，从而能够测量极小的力——例如生物细胞附着在表面或者其他细胞上的力，也可以在使用激光处理细胞(“细胞手术”)。时，用光学镊子将细胞精确地固定在适当位置。除此之外，光学镊子还可以和其他的测试方法综合应用，例如和气溶胶的光散射或者光谱学相结合，可以用来发现可能造成大气污染的颗粒。

光的机械力可以用来发现微小物体运动的全新状态。如图31所示的微型机械悬臂，其悬臂的运动就可以用光进行观察和控制。这种光力可以用来加热或者冷却悬臂的振动，就好像为机械手表的弹簧上弦

或者放松一样，最终尽可能使其达到最安静的静止状态，只有运动量子的涨落才能干扰这种完全静止状态。光力也可以用来冷却比机械悬臂小得多的原子，并揭示出物质更奇特的量子态。

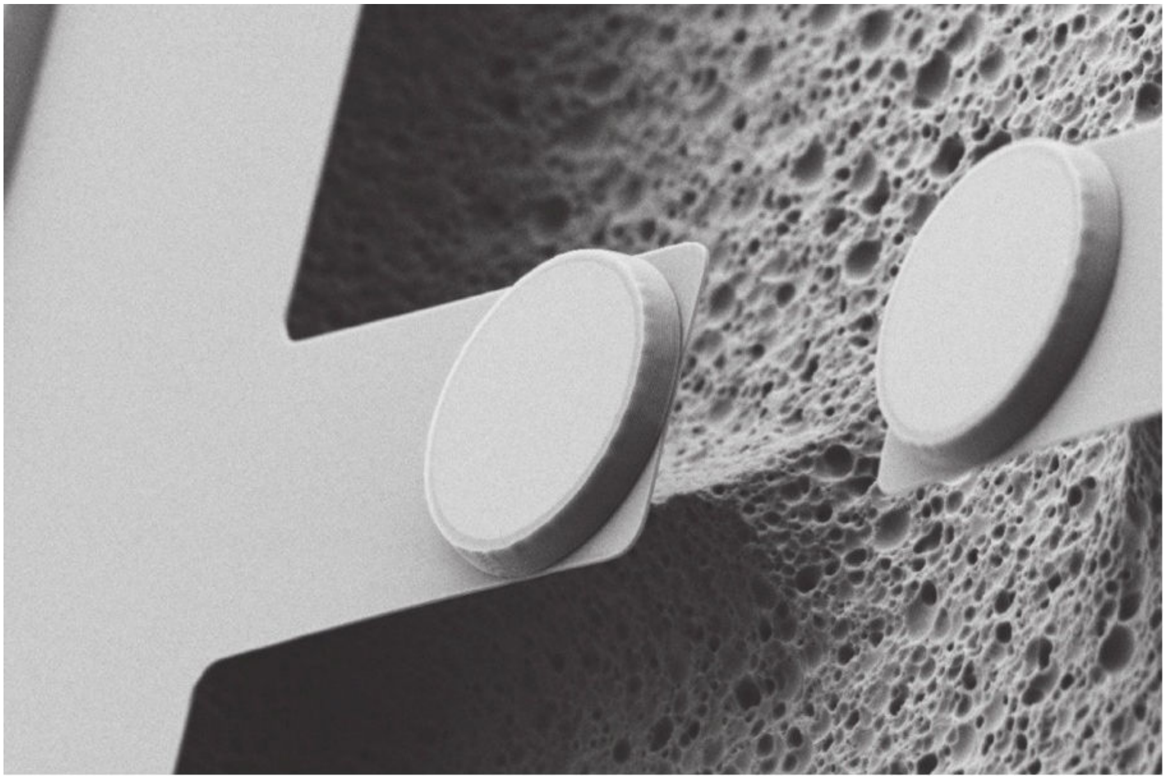


图31　一种由光的机械力控制的纳米级悬梁臂。图中的圆盘是直径约为30微米的微小镜面

超冷

你体验过的最冷温度是多少？比冬天的牛津(大约2℃)、渥太华(-20℃)或南极(-50℃)还要冷吗？或许是液氮的温度(-200℃)？这些当然都很冷，但绝不是最冷的情况。研究发现，温度有一个最低极限，-273℃或者0开尔文(单位：K)。我们将其称为“绝对零度”，没有比这个温度再低的了。要达到这个温度需要物体保持绝对静止，此

时，物体只受到量子力学的影响，使得其原子和分子产生轻微的抖动。

事实上，制造一台能达到绝对零度的机器是不可能的，但是使用“光学冰箱”，也只可以达到非常接近绝对零度的温度。当温度降低到一定程度时，原子几乎停止运动，这意味着原子的尺寸变大了(量子力学告诉我们，不能同时确定物体的精确位置和速度。如果原子完全停止运动，就意味着它必须向整个空间扩展)。因此，所有冷却的原子占据了空间中相同的区域，从而产生了一些非常奇特的新现象。

光学冰箱的工作原理是利用激光来“冷却”原子。想象有一束激光照射在一个从左向右移动的原子上，由于激光从右向左照射，所以会有一束光子直接撞击到原子上。为了使这些光子可以被原子吸收，激光根据原子运动的速度，被调节到某个特定频率。当原子从激光中吸收了一个光子时，就好像被光子当面“踢”了一下，从而降低了运动速度(更确切地说，光子的动量转移到了原子上。由于两者的初始动量方向相反，所以原子的动量减小，因此速度降低)。在之后的某个时刻原子必须重新发射光子，并且会受到与发射光子方向相反的力。但是由于原子发射光子的方向是随机的，这就意味着原子受力的方向也是随机的，因此原子可以向任意方向运动。

当你观察了足够多的吸收和散射过程，你就会发现，尽管光总是从同一个方向(激光束的入射方向)被吸收，但是原子却向各个方向发射光子，而不会偏向某个方向。这个现象导致的结果是，与入射激光束相反方向运动的一组原子会慢慢停止运动，此后开始向各个方向随机运动。这样的随机运动对应着某一温度，这一温度与原子吸收光子到重新发射出去的时长成正比。

在这个方法的基础上有几个改良版本，每一种版本方法都是利用光将原子(和分子)冷却到更低的温度。在这种情况下，光就像一种

“黏性流体”，其中的原子运动得越来越慢。一旦原子的速度降低到一定程度，甚至可以用光学镊子来捕获原子，此时就可以应用更加复杂的光学冷却技术，使温度降低到仅比绝对零度高十亿分之一摄氏度。

我之前提到，即使在绝对零度下原子依然会有“抖动”，这种抖动是量子力学引起的。可以将这种抖动的区域看作原子本身的空间区域。也就是说，根据量子力学，原子不仅以随机的方式在一个狭小的空间区域内游荡，还存在于整个空间区域内。对于被困在如此低温下的原子而言，该区域的大小可能是千分之一米。不过考虑到电子到原子核的距离只有十亿分之一米，那么这个原子所占的空间可以说是相当大了。更奇怪的是，几个原子可以同时占据同一空间区域。

这个概念与我们的直觉相悖。我们常常认为原子就像一个个小小的台球，可以紧密地堆积在一起，就像组成固体材料中的元件结构一样；同时，由于原子在材料中的位置不同，它们各自又保持着独特的特性。但这些超低温的原子就不一样了，它们可以同时存在于任何地方，这是一种新的物质状态。这种状态由爱因斯坦和印度科学家萨特廷德拉·纳特·玻色[2](Satyendra Nath Bose)发现，被称为玻色-爱因斯坦凝聚态(Bose-Einstein condensate)。

这种奇特的状态有一些显著的特性。例如，原子云[3]就是一种超流体[4]，在流动时没有黏性。此外，原子云可以被一分为二然后重组在一起，从而使这两个被分离的原子云表现出量子干涉效应。这从本质上展示了一个大物体(包含许多原子，使其大小可见)所具有的量子特性。这种量子特性可以归因于每一个特定原子的不确定性——不确定它处于这一半还是另一半原子云中。我们必须考虑每个原子其实同时存在于这两个原子云中。

由于这些冷原子可以被光束束缚，所以也可以用几束光束创造出某种空间结构来操纵原子。例如，当两个光束相遇时，会形成一个干涉图样(见第3章)，在干涉图样中有些区域强度高，有些区域强度低。冷原子们要不就都停留在强度高的区域，要不就都停留在强度低的区域(可以通过选择特定波长的光来调整原子所在的区域)。随着光束强度的增大，原子会落入干涉图样中出现的“鸡蛋托盒”状的光陷阱中(见图32a)。它们落入的方式也很有趣。

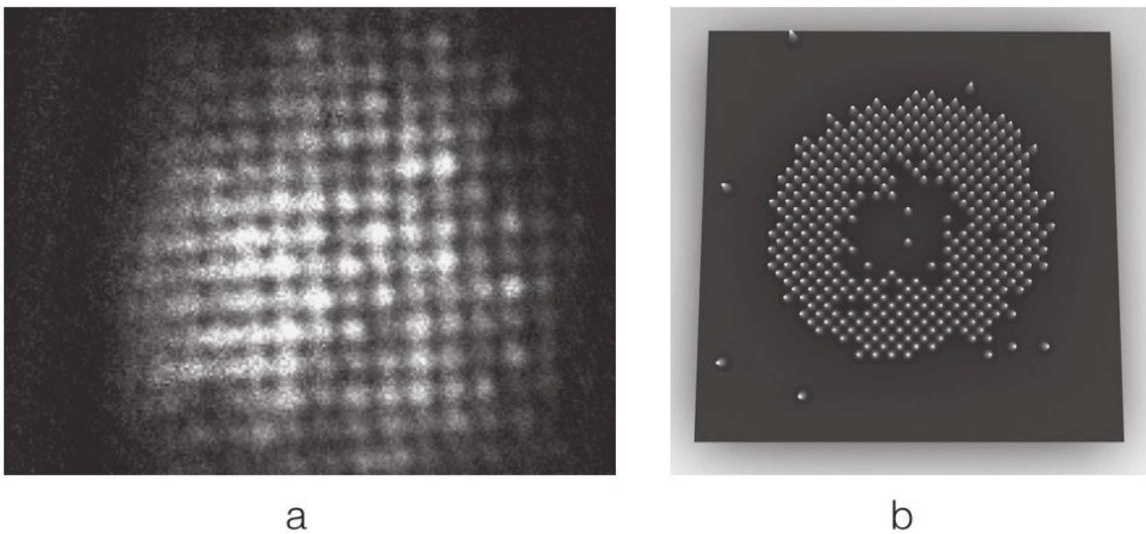


图32　被光束缚在光学晶格[5]中的冷原子：a. 光学晶格中每个晶胞束缚有几百个原子(温度为几十微开尔文)；b. 一些分布在各个“格子”中的单个原子(温度为几纳开尔文)

当原子足够冷时，它们并不喜欢待在鸡蛋托盒内的同一个“格子”里，因此最终原子的分布特别像一个完整的鸡蛋托盒——一个原子待在一个格子里，如图32b所示。这种情况下不存在超流体，因为原子们喜欢待在原地。事实上，这更像是一个“绝缘体”，因为所有的原子都没有动。通过调节光的强度，可以探究原子从完全自由流动到完全不流动的有趣过渡。

在量子力学的环境中实现对原子的控制，这使得科学家们能够探究与其他类型的材料(例如固态金属氧化物)相关的物质的新特性。尽管在这些材料上，我们很难实现同等精确度的控制和测量。现在我们可以观察到冷原子气体[6]位于“鸡蛋托盒”中的单个原子，并观察当周围环境发生变化时它们会有怎样的反应。

我们可以用许多不同类型的原子来探究这种低温状态，并且利用光来构造复杂的俘获结构。当前的一个研究领域是利用冷原子来“模拟”其他量子系统。利用这种方法可以研究其他方法无法解决的复杂问题，促进我们对材料和结构产生新的理解，从而产生新的影响。这也许可以帮助我们理解甚至是设计出新的磁铁，可用于计算机数据存储、医疗的核磁共振成像设备，甚至用于悬浮列车的无摩擦发动机。

超快

光脉冲可以非常短。在第5章中，我说过它们可以和光学场的单个周期一样短。光谱中的可见光，其光脉冲大约会持续2飞秒。对于波长较短、频率较高的极紫外(extreme ultraviolet，缩写为EUV)区，其光脉冲持续的时间更短。目前测量到的最短光脉冲的持续时间小于100阿秒(10^{-18}秒)。这些脉冲是目前可以受控产生的最短脉冲(尽管我们可以通过粒子对撞机观察到发生在更短时间尺度上的事件)。随着在X射线波段中出现的光的爆发，我们甚至可以期待产生持续时间更短的光脉冲。

这些数字简直小得让人难以置信，因此有必要选取一些参照物来帮助我们理解。宇宙的年龄是5×10^{17}秒，因此1秒和宇宙年龄的比值大约等于1阿秒和1秒的比值。或者从经济学的角度来看，如果美国的国债总量相当于1秒钟，那么1飞秒就相当于1美分。在这个尺度上，1阿秒几乎是没价值的。

在这个时间尺度上能发生什么事情呢？在第4章中我介绍了一个简单的原子模型，叫做玻尔模型。在这个模型中，电子受到电力的吸引而“环绕”原子核运动，就像行星受到万有引力的牵引绕太阳运动一样。对于简单原子(只有几个电子的原子)而言，电子绕轨道运行一周所需的时间约为150阿秒。如果我们想观察电子的运动，需要使用比150阿秒更短的光脉冲，才能使图像不致模糊。

频闪仪是与该话题最相关的一个设备。目前，研究人员已经使用了一种频闪仪的变体来观察原子和分子在基本微观层面上发生的迅速变化。在此应用中，一束激光的光脉冲被分为两个(或更多)部分，并在这两个部分之间引入一个延迟。其中，第一个脉冲先发射并照亮样品，其中一部分被样品吸收了。这“触发”了系统中的一些变化——电子在原子内部运动，或者化学键在分子或固体中振动。接着，第二个脉冲发射过去，其中一部分经由样品散射，而后被探测到。

重复进行该实验时，随着两个脉冲之间延迟时间的增加，探测到的散射光可以反映出样品的动态变化。从某种意义上来说，这是原子、分子或者固体变化的“电影”。这种“泵浦[7]探针”的方法已经被用来研究一些复杂的过程，例如在化学反应中，当两个分子通过它们之间的相互作用而被重构时发生了什么。这种方法还有一个更复杂的版本，就是利用不止两个，而是多个光脉冲。这些方法现在被用来研究许多极其有趣却令人费解的内容，从相互作用的原子、高温超导体到生物系统，等等。

我已经说过光学场的单一个周期是脉冲所能持续最短的时间。而我们可以利用由高次谐波产生的极紫外(EUV)脉冲，设计实验来测量光电场的振荡。测量脉冲场需要非常快，要比光周期快得多。我们可以利用波长更短的脉冲，它只有光学波长的二三十分之一。当一个电子被一个强光脉冲从原子上剥离时，就会产生这种波长非常短的脉冲。

这个过程需要一个光学场，其强度相当于电子与原子核之间的结合力大小。这种脉冲很容易通过在锁模激光器的输出端增加一个光放大器来获得。

当电子被强烈的脉冲从原子上剥离出来后，电子会处在一个快速振荡的电场之中。如果电子是在电场振幅为零时被剥离的，它就可以顺着光学波的下一个周期做“冲浪”运动——短暂地远离原子，然后再回来。当它返回时，速度会非常快，并且可以通过光的形式释放所有额外的能量，从而被原子重新捕获。在这种情况下，原子和电子重新结合时，会释放出非常短的脉冲。这个脉冲的波长只有几百亿分之一米，位于光谱的EUV区域，大约是产生它的光波波长的1/20。

现在想象这个EUV脉冲发射到另一个原子上。它的波长非常短，足以被原子吸收，并击出一个电子停留在原子周围。进一步设想，在同一时刻将待测量的短光脉冲也发射到这个原子上。这个脉冲的场将使停留的电子向某个方向加速，这一方向取决于电子被EUV脉冲从原子上剥离时，所处的光学波周期位置。通过改变EUV脉冲和光脉冲之间的延迟，就可以测量电子的加速度。由于速度较大的电子有更多的能量，这使我们可以“看到”光脉冲场(见图33)，尽管该场振荡的时间非常短。

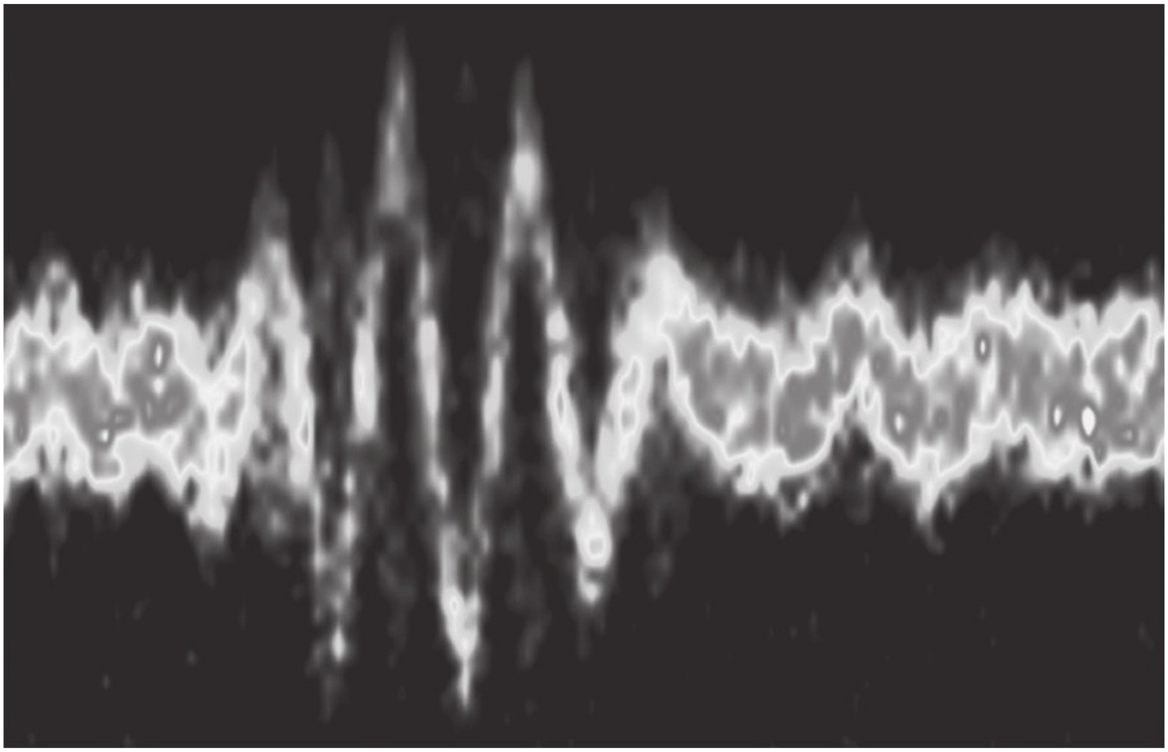


图33　光脉冲场的图像。两个相邻峰值之间的时间间隔为2.5飞秒

这种泵浦探针光谱法可以应用在生物化学的研究中，例如研究光合作用的第一步反应，即植物利用阳光的能量，将空气中的二氧化碳转化为氧气。这一步需要以极高的效率在一个大的生物分子周围传输能量。该过程包含有一些非常有趣却不能被解释的特性——其能量的传输比人们预期的要快得多，效率也要高得多。如果我们能从这种在自然界中进化了上亿年的系统中，学习如何快速高效地传输能量，或许我们就能够应用这些知识改进诸如太阳能电池的设计之类，这将给社会带来极大的影响。

超强

电费账单告诉你上个月用了多少电。它以千瓦时为计量单位，你消费的每一度电都要付费。假设你在某个月使用了220千瓦时(这是英

国每月的人均能耗)：你可以在这个月的四周内平均使用这些电，或者你也可以在第一周用光它，并在接下来的三周什么都不用。但是，你能想象在千万亿分之一秒内用完这些电吗？要在这么短的时间内用完这么多的电，你需要大量的电器，而且你还必须迅速地开关它们。因此，在这种情况下的用电峰值功率将会非常大。

光脉冲可以用完这些电。因为它们的持续时间特别短，而且包含了这么多能量。事实上，我们有可能产生某种脉冲，使其在某一时刻以相当于整个地球全部发电能力的功率提供能量。但你家里的灯并不会熄灭，这是因为脉冲太短，导致它们的总能量非常小。

产生这种脉冲需要一台巨大的激光器，其占地面积几乎赶得上一个足球场那么大。英国卢瑟福·阿普尔顿(Rutherford Appleton)实验室的火神激光器(VULCAN laser)就占地颇大。火神激光器可以让持续500飞秒的脉冲携带500焦耳(3.6×10^6焦耳 = 1千瓦时)的能量。500焦耳只是一个100瓦的灯泡在5秒内发出的能量。然而脉冲很短的持续时间意味着光的强度可以达到100万颗太阳的总量。位于加利福尼亚州利弗莫尔的美国国家点火装置(National Ignition Facility，缩写为NIF)[8]的激光器要比火神激光器大得多。还在计划中的欧洲光学基础设施(European Light Infrastructure)项目将建造一个比NIF峰值功率更大的系统。

这种非常短暂、非常强烈的光爆发可以用来改变物质的状态。光脉冲产生的最大电场大于原子内部的电子和原子核之间的引力场，所以有可能利用这个机制将原子中的电子剥离出来，从而形成一种新的物质状态——等离子体。这种剥离可以瞬间完成，比原子核运动的时间还要短，使得原子更紧密地排列在一起，因此等离子体的密度非常大，几乎与固体材料(如一块玻璃)的密度相同，然而与固体材料不同的是，等离子体的温度高达200万摄氏度。

这就是巨行星甚至一些恒星的核心所处的状态：高密度等离子体的粒子在百万倍于大气压力的压力下高速碰撞。这样的等离子体可以在实验室获得，它们的用途极为广泛。例如，我们可以通过它们了解恒星是如何工作的，它们的生命周期是怎样的，从而描述它们的进化阶段，如超新星爆炸和白矮星[9]等。其他一些令天体物理学家感兴趣的现象也可以使用激光器进行实验研究。同时，天体物理学家也使用这种等离子体来探索行星科学。例如，我们可以从气态巨行星的质量和大小来推断它们的组成，但前提是我们要知道在如此高的压力下物质能被压缩到什么程度。

一些激光设备利用波长很短的光产生脉冲。这些脉冲由磁场中加速的电子产生，所以当它们沿着加速器急速下落时，它们会左右“摇摆”。这就产生了一种由短脉冲X射线组成的同步辐射。这类激光器通常使用粒子加速器的技术甚至硬件。比如斯坦福线性对撞机光源(Linear Collider Light Source，缩写为LCLS)和德国汉堡X射线自由电子激光器(X-ray Free Electron Laser，缩写为XFEL)。

基于最强烈的激光脉冲以及X射线短时增强的技术，科学家们能够在各种条件下检测等离子体。此外，激光在原子核之间施加的巨大压力，可以在适当的条件下使原子核融合在一起，这个过程会释放出大量的能量。这种“核聚变”可能使我们获得几乎无限的能源。将激光应用于核聚变在技术上要求极高，也是目前正在探索的实现核聚变的两种方法之一：另一种方法不涉及光，光只是在其中充当一种监测工具。不过这两种方法都使用了致密等离子体。

当激光脉冲穿过等离子体时，会产生一种波，类似于船只划过水面时会留下波纹。在等离子体的电场内距离10-6米的两端电压可达到10万伏以上(这意味着约为高压电塔电线中10倍的电压，被加在只有人类头发丝直径十分之一的宽度上)。这样的电场强度比世界上用于研究

基本粒子的最大机器，如日内瓦欧洲核子研究中心的大型强子对撞机(Large Hadron Collider，缩写为LHC)所使用的加速场强度至少大1000倍。在未来，也许我们可以用激光制造出可以放在桌面上的小型设备，使电子能够加速到目前LHC所能达到的能量。

通过激光脉冲与物质相互作用产生的超强电场，也可以给较重的粒子加速，比如质子。目前质子束正作为一种治疗癌症的手段被研发，与目前使用的其他类型的放射治疗相比，重的粒子可以更精确、更深入地送达病变组织。

光的非凡特性使人们能够不断在广阔的范围里开创新领域。可以说，光是科学技术中一种无处不在的工具。

[1] 分子马达(molecular motor)是由生物大分子构成的，利用生物化学能进行机械做功的纳米系统。生命体的一切活动都依赖于分子马达。

[2] 玻色(1894—1974)，印度物理学家，奠定了玻色-爱因斯坦凝聚理论的基础。玻色子就是以他的名字命名的。

[3] 由大量原子构成，处于玻色-爱因斯坦凝聚态。

[4] 超流体是在超低温下的一种内部完全缺乏黏性的理想流体。

[5] 即文中描述的由多个光束相互干涉形成的光陷阱，可将冷原子“囚禁”于其中。

[6] 由单个原子组成的气体。

[7] 利用光将原子或者分子内部的电子从低能态激发到高能态的过程。

[8] 美国科学家制造的世界上最大的激光核聚变装置，能产生恒星内核温度和压力，被称为“人造小太阳”。

[9] 超新星爆炸和白矮星都是恒星进化到末期可能出现的结果。

08 量子光学 Quantum Light

在第1章中，我介绍了光可以被视为粒子流，为方便起见，将其记为“光子”。事实证明，光子是真实的粒子，它可以被产生、研究、测量、储存与使用。尽管从某种意义上来说，光子用最简单的方式描述了光，但是制造单个光子却并不简单。大多数光源产生不同种类的光，其中包含的光子数量是不固定的。

例如，一个灯泡会产生一股向任意方向发射的光子流。如果你从某个方向观察灯泡发出的光，仅观察一小段时间——你可以称之为一个时间间隙，那么你可以数出在这段时间内发射出的光子数。但是，如果你把这个实验重复几次，你会发现每次数出来的光子数量都是随机的，有的时候多，有的时候少。光子的平均数量其实是固定的，只取决于灯泡的亮度。但是你却永远无法肯定，在给定的时间内，能从光束中测量出多少个光子。这就是“经典光”的特征之一，即光可以完全用波来描述。

激光也是这样。激光脉冲中包含的平均光子数可以很多，但对于任何给定的脉冲，实际测得的光子数都将大于或小于平均数。一个脉冲中光子数的变动范围近似于平均光子数的平方根。而每个脉冲中光子数的变动与所有脉冲中平均光子数的比，被称为相对“噪声”，所以该“噪声”将随着平均光子数的增加而减小。

激光束有固有的强度噪声，这就限制了利用激光得到的图像质量。激光强度的波动意味着不能精确探测图像中两点间的距离。事实上，用低强度光获取的图像是非常不精确的，因为低强度光中的平均光子数很小(所以噪声大，导致物体很难被看清)，而且光子数在帧与帧之间的变化很大。获得精确结果的唯一方法是延长成像时间，从而增加到达目标物体的光子数量，再从中求取平均结果。平均信号可以

降低相对强度噪声，从而得到分辨率更高的图像，其精度与所用光子数的平方根成正比。因为经典光束无法超越这一精度，所以该精度被称为“标准量子极限”。

另一方面，在平均光子数相同的情况下，利用量子光可以获得更好的平均效果，这是因为量子光的噪声远比经典光的低。但首先，你必须有一个量子光源。这样的光源有很多种，每种都能产生一种不同的量子光。那么接下来，先让我们来看一下使光具有原始量子态的源头——光子。

单光子

如何才能制造出单个光子呢？奥托·弗里施(Otto Frisch)在1965年想出了一个可操作性很强的方案。他的想法很简单，首先将一个原子激发到激发态(关于如何施行请参阅第5章)，然后等它降至基态。在这个过程中，原子只会发出一个光子，这是因为单个原子只能存储一个“量子”能量。你可以知道原子什么时候发射出光子，因为发射光子的动量会给原子一个反向作用力，只要探测到了原子的运动，你就能确定此时原子发射出了光子，而且还可以根据原子的运动方向来确定光子的发射方向。

一些现代量子光源的工作原理与此类似，只是它们将原子困在两个平面镜(类似于激光器中的光学“腔”)之间，并快速激发原子，使其优先向沿着腔轴的方向发射光子，这就构成了一个稳定的单光子光源。由于光子的发射具有严格的规律性，因而这种单光子光源是一种特别的“低噪声”源。如果在某个特定的时间段内观察这样一束光，你可以准确预测其中有多少个光子——只有一个。因此，这种光的强度异常稳定，与“嘈杂”的经典光束不同，它是一种很“安静”的光束。

此想法也被用于其他量子光源。你可以利用非线性光学效应构造一个非常简单的光源。具体来说，某些晶体可以使一个高能量光子分裂成两个能量较低的光子，每个光子大约是原始光子的一半。对于大多数材料来说，发生这种裂变的概率相当小。由于光子成对产生，你可以将其中的一个作为“前驱体”，作为表示另一个光子存在的信号(见图34)。这类光源是量子光学领域的主力军，它利用光的量子力学特性来探索量子物理学的基础，进而发展出一些新型信息技术。

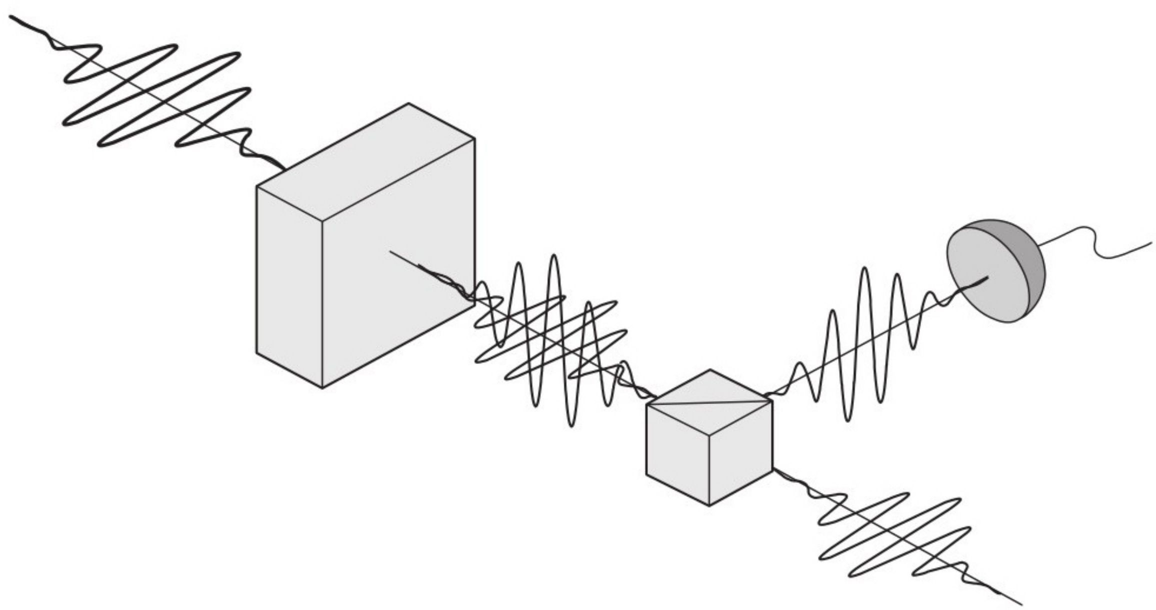


图34　“前驱体”单光子光源随机产生光子，每当一个光子产生时，“前驱体”会发出信号

正如经典的电磁波可以发生极化[1]一样，光子也可以发生偏振，因此，我们可以找到一个垂直偏振(用V表示)或水平偏振(用H表示)的光子。它们就像波一样，当我们通过观察光子能否通过水平方向的偏振器来测量其偏振方向时，会发现水平偏振的光子总是可以通过偏振器，而垂直偏振的光子总会被挡住，无法通过。

我们可以产生一个对角向偏振的光子，使其振荡角度与水平和垂直方向均呈45°。当我们尝试观察这种对角向偏振光子能否通过水平偏振器时，其结果就存在不确定性，这非常奇怪。光子是光的最小“组分”，所以它不能再被分割。那么对角向偏振光子在通过水平偏振器时会发生什么呢？实际上，对角向偏振的光子有一半的概率通过偏振器，以一半的概率被反射回来(见图35)。

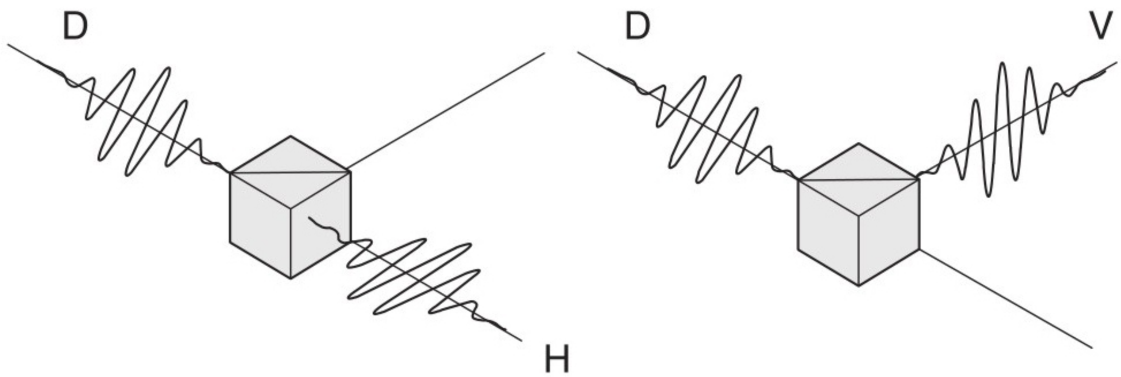


图35　一个对角向偏振(用D表示)的光子遇到一个偏振器，光子会随机地通过水平偏振器(用H表示)，或者通过垂直偏振器(用V表示)

这意味着如果你让对角向偏振(用D表示)的单光子通过水平方向的偏振器100万次，那么其中有50万次可以通过，而剩余的50万次则不能。量子力学非常奇怪的一点是，你无法预测每次实验的准确结果。这并不是因为光子有时是水平偏振的，有时是垂直偏振的，而是因为同一个光子既是水平又是垂直偏振的。因此，此随机结果从宇宙最基本的层面揭示了量子物理学所描述的内在不确定性。

在这种情况下，你可以用单光子做一些普通光不能实现的实验。例如，可以用光子的这种属性来生成随机数，方法是记录光子通过偏振器(标记为1)，还是被偏振器反射(标记为0)。由0和1组成的字符串中的随机性是由基础物理中的固有特性保证的，而不只是来自投掷骰

子或其他偶然事件中。因此，量子随机数生成器是一项新兴的业务，它们生成的随机性是无法被伪造的。

你也可以利用物理定律来保障通信链路的安全，而不是依赖电信供应商。这是基于光子的两个重要特性：第一，你不能同时在两个地方探测到同一个光子。因此，如果一个窃听者想通过捕获光子来获取你发送的信息，那么，你的信息不会被发出去，也就收不到任何回应，这个时候你就会意识到通信出了问题。当然，窃听者可能会很狡猾地给你发送一个“诱饵”光子，希望你能将这个含有伪造信息的光子当做你本该收到的光子。但是基于光子的第二个重要特性，你可以看出它是伪造的。因为在量子力学中，没有任何测量能够测得单个量子的所有信息。

结果示范如下，比如你想通过链路发送一个简单的二进制消息(由0和1组成)，并将一个垂直偏振光子标记为0，一个对角向偏振光子标记为1。如果偷听者(通常被称为Eve)检测发现光子是“垂直偏振”的，那么她依然不能确定光子是不是0，因为对角向偏振光子至少有一半概率也会给出“垂直偏振”的测量结果。所以Eve只能得到一部分信息，而不是全部。

现在，假设消息的发送者(通常称为Alice，而你是接收者，可称为Bob)向你发送一个编码为1的光子。假设Eve测量该光子是否垂直偏振，并发现它通过了检测。为了伪装，她必须选择给你发送一个垂直偏振光子还是对角向偏振光子。她最好向你发送垂直偏振光子，因为这是根据测量结果得出的最有可能的情况。而作为接收者的你可以对接收到的光子进行对角向偏振测量。由于你知道Alice发送的部分信息，因此如果光子来自Eve，那么这个光子在50%的时间里会给出错误的结果。如果它来自Alice，你永远不会得到错误的结果。因此，通过

比较接收到的信息与已知的Alice发送的部分信息，你就可以判断Eve是否篡改了你的传输链路。

然而，Eve可能想出更聪明的办法。她也许试图从Alice那里复制光子而不测量它。她可以先复制两份，并把原始光子发送给你。然后她就可以在其中一个副本上进行垂直偏振测量，在另一个副本上进行对角向偏振测量，这样就可以在你毫不知情的情况下确定Alice发送给你的光子“比特”中的全部信息。然而，即使如此她的伎俩也还是无法得逞。因为量子力学的一个显著特点是，不可能制造出一台复印机能精确复制或克隆一个处于未知量子状态的单个粒子，这是违背物理定律的。由于物理上的这两个限制(“不可测量”和“不可克隆”)，我们可以在Alice和你之间建立一个安全的通信链路，用来传输秘密的随机比特流。

压缩光

其他种类的量子光相比经典光也有一些性能上的提升。我们知道，光是电磁场的振荡。几乎可以将激光束视作这种振荡的理想状态。然而，即使是激光，它的振幅中也有一些“噪声”。也就是说，每次测得的振幅，并不完全一样。如图36a，图中显示了每一点或每一相位的场的振荡不确定性。有一种特殊的量子光被称为“压缩光”，这种光的噪声会随着光场周期的相位而变化，如图36b所示。它在某些相位的噪声比其他相位的更大。结果表明，这样的场只能由成对的光子生成。如果你测量光子的数量，你只能得到偶数的结果。正是这些光子对量子的干涉形成了与相位相关的振幅噪声。

这一特性自有其应用价值。假设你想测量波的相位。回想一下，之前介绍过使用干涉仪时，光束的相移是由被测量的物体(比如某个特定分子)引发的。在光场波动最小的地方可以更精确地测量波的相位。事实上，在某些相位，压缩光场的起伏比任何经典场的都要小，因此

使用压缩光场的相位传感器将比使用经典光场的传感器更精确。它们甚至可以打破标准量子极限。

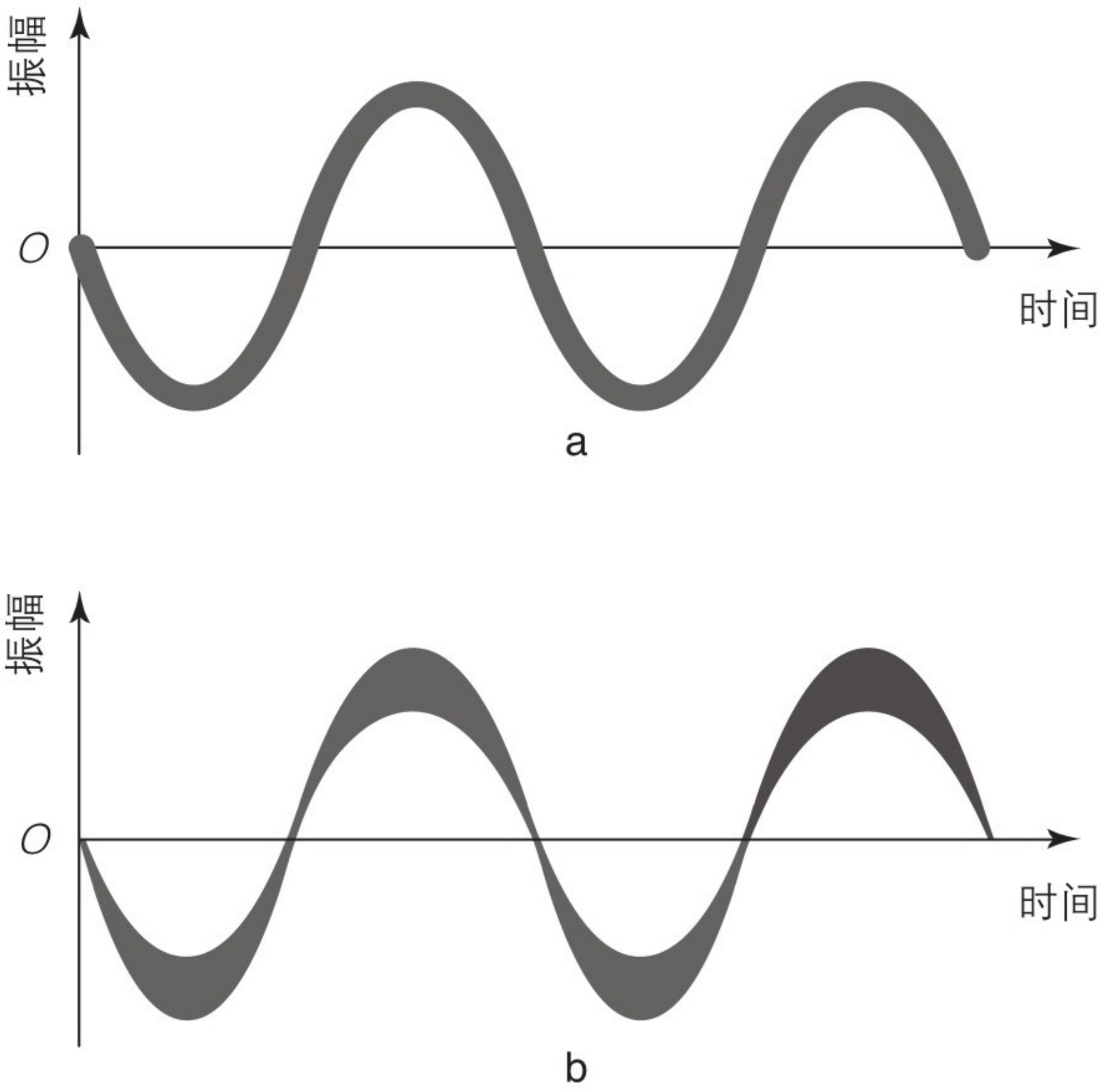


图36　压缩光a与激光b相比，a在其振荡的某些点上的场振幅噪声减小了

目前，这种测量方法仍然比较昂贵，因此只有在比其他方法具有明显优势时才会使用。例如，用大型光学干涉仪探测引力波，像德国汉诺威附近的GEO 600项目，就需要使用压缩光检测相移，而相移的变化又与引力波引起的光的相对路径距离的变化相对应。因此，可以通

过测量相移的变化来判断引力波是否存在。当然，此时光的相对路径距离的变化非常小，如果与地球到太阳的距离相比，这段距离只有一个原子大小。

量子纠缠

当存在一个以上的量子光束时，事情会变得更加奇怪。光子会以某种方式缠绕在一起，导致我们无法区分任何一束光的属性，例如颜色、位置、方向或脉冲形状。这远远超出了波粒二象性的基本概念。它挑战了经典世界的观念，即在经典世界中可以确定物理实体的某种属性(如光束的频率、到达时间或水平/垂直偏振等等)。这些属性可以被测量，并且彼此自洽。但是，这并不适用于在某种状态下的成对光束，而这一点可以通过实验来验证。这是20世纪基础光学的伟大成就之一。

借助这一特性，我们可以利用量子光学来检验伟大的爱因斯坦、鲍里斯·波多尔斯基[2](Boris Podolsky)和内森·罗森[3](Nathan Rosen)的著名猜想，即判断他们对于粒子系统的量子力学描述是不是完整的，而且不需要借助任何其他信息就可以确定该系统的方方面面。约翰·贝尔[4](John Bell)在20世纪60年代找到了一种量化这类问题的方法，并尽力尝试用实验来检验他的假设。这些实验一般被称为“贝尔实验”，其中最早实现也是目前最具说服力的研究中就使用了光子对，并且每对光子之间都彼此相关。正是这些相关性，使得量子粒子与经典粒子大为不同。为了更全面地了解这种量子效应的奇特性，我们有必要在此进行更深入的探讨。

相关性几乎无处不在。以下面这个简单的游戏为例。发牌人拿两副牌，一副牌的背面是绿色，另一副牌的背面是蓝色。发牌人从每副牌中各挑一张，其中一张牌给你，另一张给另一位玩家。当然，你们

的牌背面的颜色总是不同的，但它们正面的花色却可能相同，因为每副牌中有一半正面是黑色花色，另一半是红色花色。事实上，你会认为在一半的时间里，两个人拿到的花色应该是相同的，这是因为出现红色红色和黑色黑色的概率为50%。

如果你和你的搭档每次都同时拿到红色或黑色的花色，这些卡片就被认为是“相关的”。这是你所能想象到的最强烈的相关。事实上，如果你们两个人在超过一半的时间里都得到了相同花色的牌，也可以说这些卡片是相关的，尽管这种相关性会比第一个例子中的“弱”。通过测量相关性，你可以确定发牌人是否在作弊，因为我们一般认为发牌人所用的是两副独立、完整的牌。

我们可以用偏振代替牌来解释光子的这种相关性。也就是说，一个水平偏振的光子可能被看作“红色”光子，而一个垂直偏振的光子可能被看作“黑色”光子。如果一个光源一次产生两个光子，就像之前描述的，它总是产生具有特定偏振方向的光子，比如一个垂直的和一个水平的，或者两个都是水平的，就可以说这个光源产生了相关的光子束。这种类型的相关性被称为**经典相关**，因为它和经典对象(如扑克牌)的情况一样。

相关性在其本质上具有量子力学的特性。假设一对光子对有两种可能的初始状态，第一种是水平偏振，第二种是垂直偏振，或者第一种是垂直偏振，第二种是水平偏振。在经典世界中，两个粒子的这两种偏振组合是相互排斥的：也就是说这两个光子要么是水平-垂直偏振组合，要么是垂直-水平偏振组合，并且每一种组合发生的概率都是50%。但是，正如单光子可以同时处于水平偏振和垂直偏振的叠加态一样，这对光子也可以处于叠加态，即水平-垂直偏振和垂直-水平偏振的叠加态，如图37所示。这种相关性比任何经典粒子更强，被称为纠缠。它是量子物理学中最神奇的性质，并产生了深远的影响。

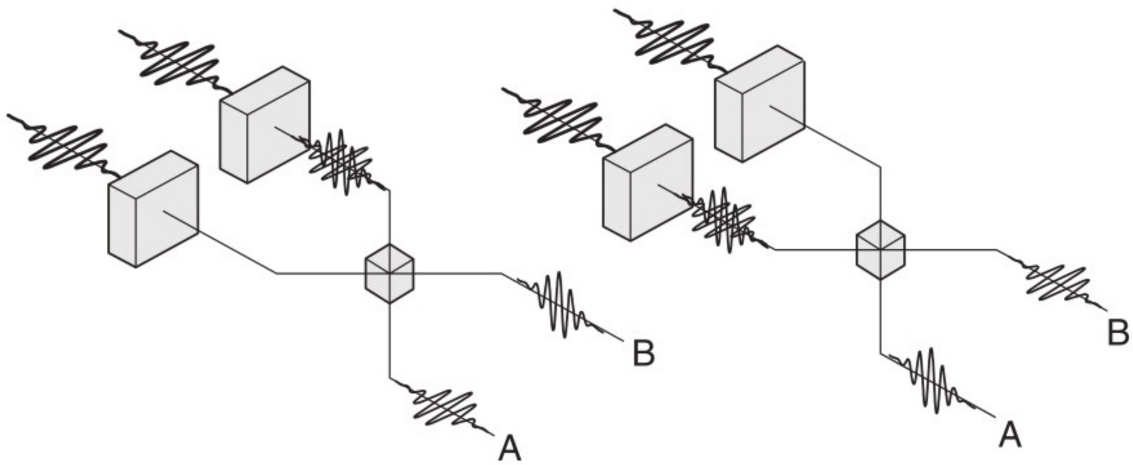


图37　产生偏振纠缠光子的光源

这些都是贝尔实验所揭示出来的。在这样的测试中，你不仅要考虑每个粒子的水平偏振和垂直偏振之间相关的可能性，还要考虑对角向偏振(用D表示)和反对角向偏振(用A表示)之间相关的可能性，这两个偏振的方向角度都在水平和垂直偏振的正中间(对角向偏振光的一个例子展示在图35中。反对角向偏振的方向与对角向偏振的成直角)。用纸牌来类比的话，也就是说你看到的纸牌花色不是红色(相当于水平偏振)就是黑色(相当于垂直偏振)。或者你可以看到纸牌的背面，不是绿色(相当于对角向偏振)就是蓝色(相当于反对角向偏振)。

一个量子游戏

现在想象一个纸牌游戏，发牌人从任意一副牌中抽牌并发给每个玩家一张牌。这意味着，每个玩家手中有一张牌的正面花色(用“正”表示)为红色(用R表示)或黑色(用B表示)，牌的背面颜色(用“背”表示)为绿色(用g表示)或蓝色(用b表示)。发牌人选择用以下的方式发牌：如果一个玩家只能看到他自己牌的正面，另一个玩家只看到他自己牌的背面，用(正，背)表示，那么在这种情况下，两位玩家间永远不会出现(R，　B)[5]的结果。同样，如果第一个玩家看他自己牌的背

面，另一个看他自己牌的正面，即(背，正)，那么他们也永远不会看到(B， R)的结果。然而，当两个玩家都看的是牌的正面(正，正)，他们有时会看到(R， R)。由此，你可以根据逻辑推理得知，如果他们都看牌的背面(背，背)，他们就一定会看到(g， g)。这就是经典的纸牌游戏中会发生的情况。

但事实上，当你用量子相关的光子(或其他粒子)做这样的游戏时，结果将完全不同。当第一个玩家测量水平偏振、第二个玩家测量对角向偏振(或者两个玩家调换)时，结果发现他们永远也得不到(V = 0， D = 1)和(D = 1， V = 0)这两个结果。而当他们都使用对角向偏振进行测量时，他们有时会得到结果(D = 1， D = 1)。如果据此进行逻辑推理，当他们都使用水平方向的偏振片测量光子时，他们应该有时会得到结果(H = 1， H = 1)，但是，当他们真正做这个实验时，他们却永远不会得到这个结果。这种量子卡牌游戏的所有可能结果都在表1中。这样的实验是可操作的，而且已经使用光子验证过了。

测量方式	测量结果：玩家 1；玩家 2；概率			
正，正	R; R; 1/2	R; B; 1/12	B; R; 1/12	B; B; 3/4
正，背	R; b; 0	R; g; 1/6	B; b; 2/3	B; g; 1/6
背，正	b; R; 0	b; B; 2/3	g; R; 1/6	g; B; 1/6
背，背	b; b; 1/3	b; g; 1/3	g; b; 1/3	g; g; 0

表1　一个量子纸牌游戏可能出现的结果概率表

事物的定域属性

那么到底发生了什么呢？这是量子物理学一个根本的怪异之处：量子卡牌游戏的结论是，光源产生光子时，光子的偏振并没有被预先确定。这就好比从一副卡牌中取出背面颜色未定的牌。这与我们的经验相悖，每张纸牌的正面有一定有特定花色，背面一定有特定颜色。无论我们是否清楚，甚至是发牌人是否清楚花色或者颜色是什么，我们都会非常确信，当这些牌被发给我们时，它们确实具有特定的花色和颜色。我们当然不希望它们的这些属性被我们的行为影响。但是量子力学却告诉我们，我们不能预先确定纸牌的花色和颜色。

只有通过测量才能给出确定的结果。但是我们不能简单地认为，测量出来的结果就是之前不知道的、预先设置的光子特性。事实上，当光源产生单个光子时，你不能事先确定光子的偏振，因而实际测量的结果并不是其事先具有的偏振，而是如果没有测量，你将永远不知道其偏振状态。用纸牌游戏来说，如果你企图设计出一种发牌方式得到相同的效果，你会发现那是不可能的，除非这些纸牌可以同时以红黑或绿蓝的方式叠加[6]。对光子来说也是一样，它们必须处在水平和垂直偏振的叠加态，形成某种特定的相关性，这种相关性被称为“量子纠缠”。

纠缠是一个非常奇怪的概念，我们不可能从普通日常的角度来考虑它，例如之前举的纸牌游戏的例子。然而，纠缠也很常见，它出现在许多量子尺度的事物中，甚至也发生在日常生活中。例如，分子中电子之间的相关性就是纠缠；在构成分子的原子之间，甚至是相对较小的原子之间也可以通过纠缠形成化学键；纠缠还造就了像超导体这样的特殊物质。

令人惊讶的是，纠缠也可以被运用到现实技术中。你很难想象这样一个神秘而抽象的概念可以有任何实际用途，但它确实有用。它使

很多信息处理的方法得以实现，而这些方法是不能够通过来回发送经典波来实现的。

事实上，所有的信息处理系统都是建立在物质实体之上，这表明，这些系统必定反映了其组成部分的物理原理(通常是经典物理)。这使得人们看到了新的机遇，即可以基于量子力学来建造新的系统，如在计算、通信和测量技术方面创造新技术，并以前所未有的方式超越当代技术。例如，未来通信的安全性将由自然法则所保障；未来的计算机能够解决现在我们根本“无法计算”的问题；未来的成像系统可以让我们看到从未看到过的物体。

光在创建这样的系统中起着重要的作用。例如，光纤网络的基础设施可以用来在双方之间绝对安全地发放随机的“量子密钥”(0和1的随机字符串)，然后再利用“量子密钥”来编码需要传递的消息。这样的网络还可以用来连接小型量子处理器，最终组成分布式量子计算机。事实上，研究也表明，原则上可以完全不借助光来建造量子计算机，尽管这种方法极具挑战性。这些技术的结合为量子互联网的未来带来了希望，这是一种与我们目前使用的技术完全不同的通信和处理信息的方式，而且都将通过光来实现。

[1] 改变电磁波电场强度的方向等性质，光学上被称为偏振。

[2] 美国物理学家，为了论证量子力学的不完备性，于1993年与爱因斯坦和内森•罗森一起提出了EPR佯谬。

[3] 以色列物理学家，与爱因斯坦和鲍里斯•波多尔斯基一起提出了EPR佯谬。

[4] 爱尔兰物理学家，发展了量子力学中重要的贝尔定理。

[5] 原文为(R， b)，意为红色蓝色可以出现，疑为出现错误。

[6] 即纸牌处在一种状态下，使得它的正面花色同时既是红色也是黑色，背面颜色既是蓝色也是绿色。

09 黎明时分 Twilight

德谟克利特的原子

和牛顿的光粒子

是红海海岸上的沙粒

在那里以色列的帐篷竟是如此闪耀。

——威廉·布莱克(William Blacke)

《 嘲笑吧，嘲笑吧，伏尔泰，卢梭》

光，有着如此悠长的历史与广泛的应用，因此这本薄薄的书必定无法涵盖关于它的所有话题。尤其是那些在可见光区域之外的电磁光谱上观察到的令人惊叹的发现，以及日常生活中随处可见的光学仪器，这些都没有在书中得到详尽的描述。

从多镜面望远镜(直径可达几十米，其自适应控制的反射镜面可以抵消由大气湍流引起的空中观测目标的闪烁现象)到大规模同步加速器(对电子加速直到它们辐射强烈的X射线，可以用来观察自然或人造材料的结构，从而揭示例如生物重要分子或工程金属的结构)，我们已经拥有了无数各式各样的成像设备。无论如何，我希望你已经感受到了光之美，知道了人类为揭开光的神秘面纱所经历的旅程，也了解了各种光之谜是如何被一一解开的。

与此同时，光学技术的蓬勃发展，也开辟了新的探索领域和应用领域，并常常取得出人意料的成果。正如过去的几个世纪一样，光将持续向我们展示新的谜题，不断激发我们的想象力，并使我们研制出新的设备。

全书完

术语对照

3D films 3D电影

γ -rays γ 射线

A

Abbe, Ernst 恩斯特·阿贝

Abbe’s criterion 阿贝准则

aether 以太

Al-Hazen 海什木

aluminum 铝

Ampère, André-Marie 安德烈·马利·安培

angle of incidence 入射角

angle of reflection 反射角

angle of refraction 折射角

astronomy telescope 天文望远镜

atom 原子

attoscience 阿秒科学

aurorae 极光

B

Bacon, Roger　罗杰・培根

Balmer, Johannes　约翰・巴耳末

Bell, John　约翰・贝尔

Bell tests　贝尔实验

Betzig, Eric　埃里克・白兹格

binoculars　双筒望远镜

birefringence　双折射

birefringent materials　双折射材料

black bodies　黑体

Blake, William　威廉・布莱克

Blu-ray　蓝光

Bohr, Niels　尼尔斯・玻尔

Bose, Satyendra Nath　萨特廷德拉・纳特・玻色

Bose-Einstein condensate　玻色 - 爱因斯坦凝聚态

Brewster, David　大卫・布鲁斯特

brightness　亮度

Broglie, Louis de　路易斯·德布罗意

C

caesium　铯

calcite　方解石

chromatic aberration　色差

classical correlations　经典相关

classical light　经典光

coherence　相干性

compact disc　光碟(CD)

constructive　相长

contact lenses　隐形眼镜

copper　铜

cornea　角膜

correlations　相关性

Coulomb, Charles Augustin de　查尔斯·奥古斯丁·德·库仑

Crick, Francis　弗朗西斯·克里克

D

De Luce 《 论光》

destructive 相消

Diamond Light Source 钻石光源

diffraction 衍射

digital video disc 数字影盘(DVD)

Dirac, Paul 保罗·狄拉克

E

Edgerton, Harold 哈罗德·埃哲顿

Einstein, Albert 阿尔伯特·爱因斯坦

electric charges 电荷

electromagnetic field 电磁场

electron 电子

entanglement 纠缠

Euclid 欧几里得

Euler, Leonhart 莱昂哈德·欧拉

European Light Infrastructure 欧洲光学基础设施

extreme ultraviolet 极紫外(EUV)

F

Faraday, Michael　迈克尔·法拉第

femtochemistry　飞秒化学

Fermat, Pierre de　皮埃尔·德·费马

fibre optic broadband　光纤宽带

fiuorescent　荧光

focal length　焦距

Fourier, Joseph　约瑟夫·傅里叶

Franklin, Rosalind　罗莎琳·富兰克林

Fraunhofer, Joseph von　约瑟夫·冯·弗劳恩霍夫

frequency　频率

frequency comb　频率梳

frequency shift　频率改变

Fresnel, Auguste　奥古斯丁·菲涅耳

fringes　条纹

Frisch, Otto　奥托·弗里施

G

Gabor, Denis　丹尼斯·加伯

Galilei, Galileo　伽利略·伽利雷

global positioning system　全球定位系统(GPS)

Goethe, Johann Wolfgang von　约翰·沃尔夫冈·冯·歌德

Grimaldi, Francesco　弗朗西斯科·格里马尔迪

Grosseteste, Robert　罗伯特·格罗斯泰斯特

guide star　导星

H

Hamilton equation　哈密顿方程式

Hamilton, William Rowan　威廉·卢云·哈密顿

handedness　手性

Hell, Stefan　斯特凡·赫尔

herald　前驱体

Hertz, Heinrich　海因里希·赫兹

high-harmonic generation　高次谐波产生

hologram　全息图

Hooke, Robert　罗伯特·胡克

Hubble Space Telescope 哈勃太空望远镜

Huygens, Christian 克里斯蒂安·惠更斯

I

incoherent 不相干的

information 信息

infrared 红外线

intensity 强度

interference 干涉

irradiance 辐照度

K

Kepler, Johannes 约翰尼斯·开普勒

L

Lamb, Willis 威利斯·兰姆

Large Hadron Collider 大型强子对撞机

laser 激光或激光器

laser light pulses 激光脉冲

laser projector 激光投影仪

Laser-assisted Subepithelial 激光辅助原位角膜磨削术

Keratomileusis Leibniz, Gottfried Wilhelm von 戈特弗里德•威廉•冯•莱布尼茨

lens 透镜

light-emitting diodes 发光二极管

light sources 光源

liquid crystal 液晶

liquid crystal display 液晶显示器(LCD)

M

magnetic materials 磁性材料

magnification 放大率

Maupertuis, Pierre Louis 皮埃尔•路易•莫佩尔蒂

Maxwell, James Clerk 詹姆斯•克拉克•麦克斯韦

metamaterials 超颖材料

microscope 显微镜

microwaves 微波

mode-locked laser 锁模激光器

mode-locking 锁模法

modes 模式

molecules　分子

Muybridge, Eadweard　埃德沃德·迈布里奇

N

National Ignition Facility　美国国家点火装置

National Institute of Standards　美国国家标准与技术研究院

and Technology National Physical Laboratory　英国国家物理实验室

Newton, Isaac　艾萨克·牛顿

no-cloning　不可克隆

no-measurement　不可测量

normal materials　普通材料

O

Oersted, Hans Christian　汉斯·克里斯蒂安·奥斯特

Ohm, Georg　乔治·欧姆

optical analogy　光学类比

optical cavity　光学腔

optical clockwork　光学钟表

optical fibres　光纤

optical telecommunications　光学通信

optical tweezers　光学镊子

optics　光学

P

Pendry, John　约翰·彭德里

photo-activated localization microscopy　光激活定位显微镜

photoelectric effect　光电效应

photon　光子

Physikalisch-Technische Bundesanstalt　德国联邦物理技术研究院

pixel　像素

Planck, Max　马克斯·普朗克

plasma　等离子体

Podolsky, Boris　鲍里斯·波多尔斯基

polarization　偏振

polarizer　偏振片

protein　蛋白质

proton　质子

proton beams　质子束

pump-and-probe　泵浦探针

Q

quanta　光子

quantum entanglement　量子纠缠

quantum fields　量子场

quantum key distribution　量子密钥

quantum light　量子光

quantum random number generators　量子随机数生成器

quantum vacuum　量子真空

R

radio waves　无线电波

ray　光线

ray bundle　光束

reference wave　参考波

reflection　反射

refraction　折射

refractive index　折射率

relativity　相对论

Rosen, Nathan　内森·罗森

ruler　长度标准

Rutherford Appleton Laboratory　卢瑟福·阿普尔顿实验室

S

Schrödinger, Erwin　埃尔温·薛定谔

Smith, Adam　亚当·斯密

Snell, Willebrord　威理博·斯涅尔

space-time　时空

spectral line　光谱线

spectroscopy　光谱学

spectrum　光谱

spontaneous emission　自发辐射

squeezed light　压缩光

standard quantum limit　标准量子极限

Stanford, Leland　利兰·斯坦福

stimulated emission　受激辐射

Stimulated Emission Depletion　受激发射损耗显微技术 Microscopy

stroboscope　频闪仪

strontium　锶

super-resolution　超分辨率

synchrotron　同步加速器

synchrotron radiation　同步辐射

T

Taylor, Geoffrey　杰弗里·泰勒

time dilation　时间膨胀

transverse wave　横波

T-rays T　射线

U

ultraviolet　紫外线

V

vacuum fluctuations　真空波动

Veselago, Victor　维克托·韦谢拉戈

Volta, Allessandro　亚历山德罗·伏特

Voltaire　伏尔泰

W

warmth　温暖

Watson, James　詹姆斯·沃森

Watts　瓦特

wave　波

wave function　波函数

wave guide　波导

wavefronts　波前

wavelength　波长

wave-particle duality　波粒二象性

Wilkins, Maurice　莫里斯·威尔金斯

X

X-ray Free Electron Laser X射线自由电子激光器

X-rays X 射线

Y

Young, Thomas 托马斯·杨

Z

Zewail, Ahmed 艾哈迈德·泽维尔

OUP CORRECTED PROOF – FINAL, 07/06/2015, SPi

Ian Walmsley

LIGHT

A Very Short Introduction

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© Ian Walmsley 2015

The moral rights of the author have been asserted

First edition published in 2015

Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, by licence or under terms agreed with the appropriate reprographics rights organization. Enquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2015937261

ISBN 978–0–19–968269–0

Printed in Great Britain by
Ashford Colour Press Ltd, Gosport, Hampshire

To Kate, Alex, Nathaniel, and Clair

Contents

Preface and acknowledgements

Light enables us to see things. This tautology is the starting point of a long and fascinating trail from which developed both extraordinary views of how the natural world behaves and the very tools and methods on which modern science is built—explanations that are commensurate with experiments.

Light carries information about our surroundings, from distant stars and galaxies to the cells in our bodies to individual atoms and molecules. It is the basis of many technologies that enhance our quality of life: the Internet is powered by light; the most precise clocks in the world rely on light; the tiniest objects, from individual atoms, to live biological cells, can be observed and manipulated using light; images and displays are everywhere. Light reveals the full strangeness of the quantum world and inspires our imagination about the world. Yet it might surprise you to know that what light actually is was only really understood less than a hundred years ago, and even now we are ekeing out new insights from that understanding.

This book explores how people have come to our current view of what light is and what it does. It's a great story, reaching back into the ancient world, with a global cast of contributors, from Euclid in Athens and Al-Hazen in Baghdad, conceiving the idea of light rays, to Ted Maiman in Los

Angeles and Shuji Nakamura in Tokushima developing new lasers, by way of Joseph Fresnel in Paris and Thomas Young in London thinking about light waves, to James Clerk Maxwell in Aberdeen and Heinrich Hertz in Berlin developing from these ideas the concept of electromagnetic fields, that finally culminated in Albert Einstein in Bern and Paul Dirac in Cambridge explaining these apparently mutually incompatible facts by a radical new concept—the quantum field—that showed how it is possible to be both a particle and a wave at the same time.

At each step, new understanding (for example, of refraction) has led to new applications (such as eyeglasses for correcting vision): the path from discovery to technology is often very short when light is involved. The impact of light on the modern world is immense, and often unrealized. For this reason, 2015 has been designated by the United Nations as the International Year of Light: a celebration of light and what is made possible by it.

I should like to record my thanks to Alex Walmsley and Latha Menon for helpful comments on drafts of this book, and to many colleagues for responding to specific technical questions as well as those who kindly provided illustrations for the figures. Errors and statements that go beyond the exhortation, attributed to Einstein, to make things as simple as possible, but not simpler, are due to me.

INTERNATIONAL
YEAR OF LIGHT
2015

List of illustrations

1 George Richmond, The Creation of Light 2

? Tate, London 2015

2 Neil Armstrong’s photograph of Edwin Aldrin 5

Courtesy of NASA

3 Isaac Newton, Johann Wolfgang von Goethe, and Rosalind Franklin 9

National Portrait Gallery, London.

? Corbis; Neue Pinakothek, Munich.

? Peter Horree/Alamy; ? Science

Source/Science Photo Library

4 The spectrum of electromagnetic waves 11

5 Euclid’s construction of light rays 16

6 Rays of light reflecting off a flflat mirror and a curved mirror 18

7 The change of handedness of objects reflected in a mirror 19

8 Image formation by a ray bundle using a curved mirror 20

9 Newton’s reflecting telescope 21

? The Art Archive/Royal Society/Eileen Tweedy

10 Refraction of a ray 22

11 Image formation by a ray bundle using a lens 24

12 Descartes’ experiment to show the image formed by an eye is inverted 26Wellcome Library, London

13 Images of a flea and the nervous system of a fruit-fly larva taken using microscopes 29

Wellcome Library, London; Thomas Germe, Robert Parton and Ilan Davis, University of Oxford

14 Section of a lens used for photolithography of computer chips 31

Courtesy of Nikon Tec Corporation

15 An illustration of birefringence 35

Photographed by the author

16 A circular wave on a water surface 37

17 Interfering circular waves, constructive interference, and destructive interference 38

18 Thomas Young’s experiment 40

19 How a hologram is constructed 42

20 An X-ray diffraction pattern of a protein crystal 47

By kind permission of Dave Stuart, University of Oxford

21 Fermat’s conception of a light ray 53

22 Hamilton’s idea of rays as connecting wavefronts 56

23 Spectrum of light emitted from a neon lamp and the Sun (a ‘black body’) 59

24 An interference pattern 64

By kind permission of Markus Aspelmeyer, Universitat Wien

25 The ‘kink’ in the field of an electron 68

26 An atom undergoing absorption, spontaneous emission, and stimulated emission 72

27 A bullet frozen in motion using stroboscopic imaging 81

? 2010 MIT. Courtesy of MIT Museum

28 A scaffold representing space made up of a3D lattice of measuring rods 85

29 Time dilation due to relative motion 88

30 Train of identical nearly single-cycle optical pulses 91

31 A nano-scale cantilever 97By kind permission of Markus Aspelmeyer, Universitt Wien

32 Cold atoms trapped in an optical lattice 101

By kind permission of Chris Foot, University of Oxford; Immanuel Bloch and Stefan Kuhr, Max Planck Institut für Quantenoptik, Garching

33 Direct image of the electric field of an optical pulse 105

By kind permission of Eleftherios Goulielmakis and Ferenc Krausz, Max Planck Institut für Quantenoptik, Garching

34 A 'heralded' single photon light source 112

By kind permission of Peter Mosley, University of Bath

35 A diagonally polarized photon encounters a polarizer 113

36 Squeezed light 116

37 A light source for generating polarization entangled photons 119

38 Table of the probabilities of possible outcomes for a quantum card game 120

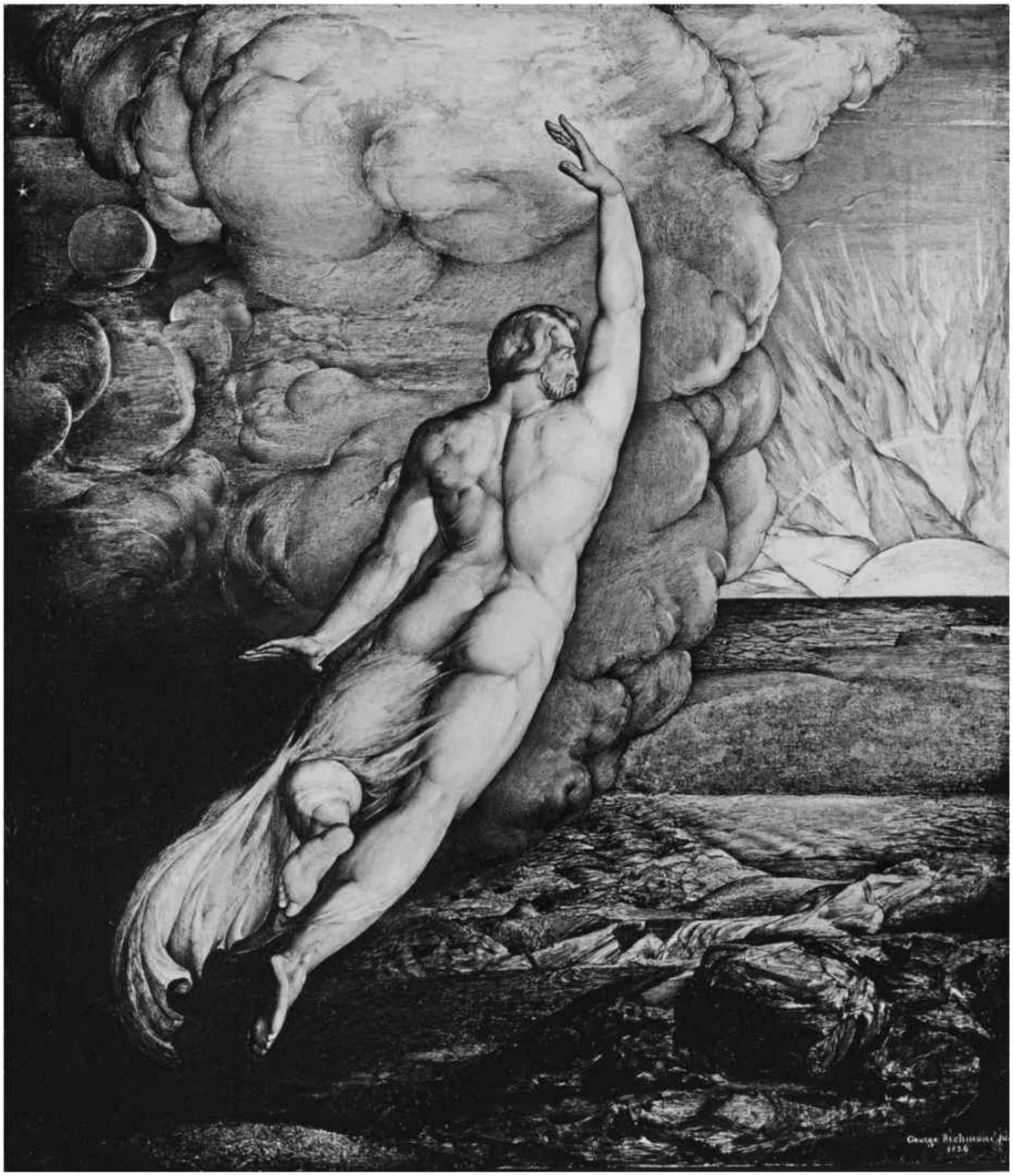
Chapter 1 What is light?

Light enables us to see the world around us. It provides the means by which our sense of sight gleans the most direct information about the physical arrangement of the world and how it is changing. Indeed, the capacity of light to carry and convey information is perhaps its most important, and remarkable, characteristic.

Seeing is believing

Sight enables us to locate ourselves in our surroundings, defining things outside ourselves that allow us to construct a true picture of the world. And sight inspires the imagination beyond the physical sensation of vision itself. George Richmond’s painting, The Creation of Light, shown in Figure 1, illustrates the central place light has in our psyche. Indeed, words deriving from the idea of light—insight, illumination, clarity, for example—pertain to human, as well as physical, qualities. In fact, Latin has two words to describe light, lux and lumen, denoting both the material and the metaphysical aspects of light. It is the former with which this book is primarily concerned. The intertwining of the physical and the poetic has made light a metaphor for thinking about the world, in philosophy, theology, psychology, art, and literature. Because it is something of which almost everyone has direct experience, the physical basis of light and how it facilitates this powerful

sense has made it an object of study for philosophers and more recently scientists, for centuries.



1. The Creation of Light by George Richmond.

Light gives life. Literally, light plays a vital role in the biological and chemical processes that underpin our and our planet's existence. Figuratively, light frames our perception of our surroundings. Our common everyday experiences illustrate the central importance of light in this regard. Of course we use it for illuminating our environment, either naturally, by the Sun or Moon, or artificially. Most common light sources use electricity, but on occasion we still use chemical reactions to generate light; burning candles, for example. The different character of the illumination has an effect on how we perceive our surroundings: it sets a 'mood' for the physical space.

Light has a very fundamental role in making possible life itself. The primary source of energy for the Earth is ultimately the Sun. And the means by which energy is transmitted from the Sun is light, both the visible components we can see, as well as some invisible ones we cannot see directly. Think, for example, of lying on a beach or sitting out in the garden in the sunshine: the warmth we feel is a consequence of some of the 'invisible light' radiating from the Sun. This is just one example of the physiological effects of light.

But at the very core of our planet's ability to sustain life is an awesome biochemical process that converts 'waste' molecules—carbon dioxide—into 'useful' ones—oxygen, using sunlight as a source of energy. The reverse process, conversion of oxygen into carbon dioxide, occurs

during respiration, as well as from the things we burn to power our everyday world.

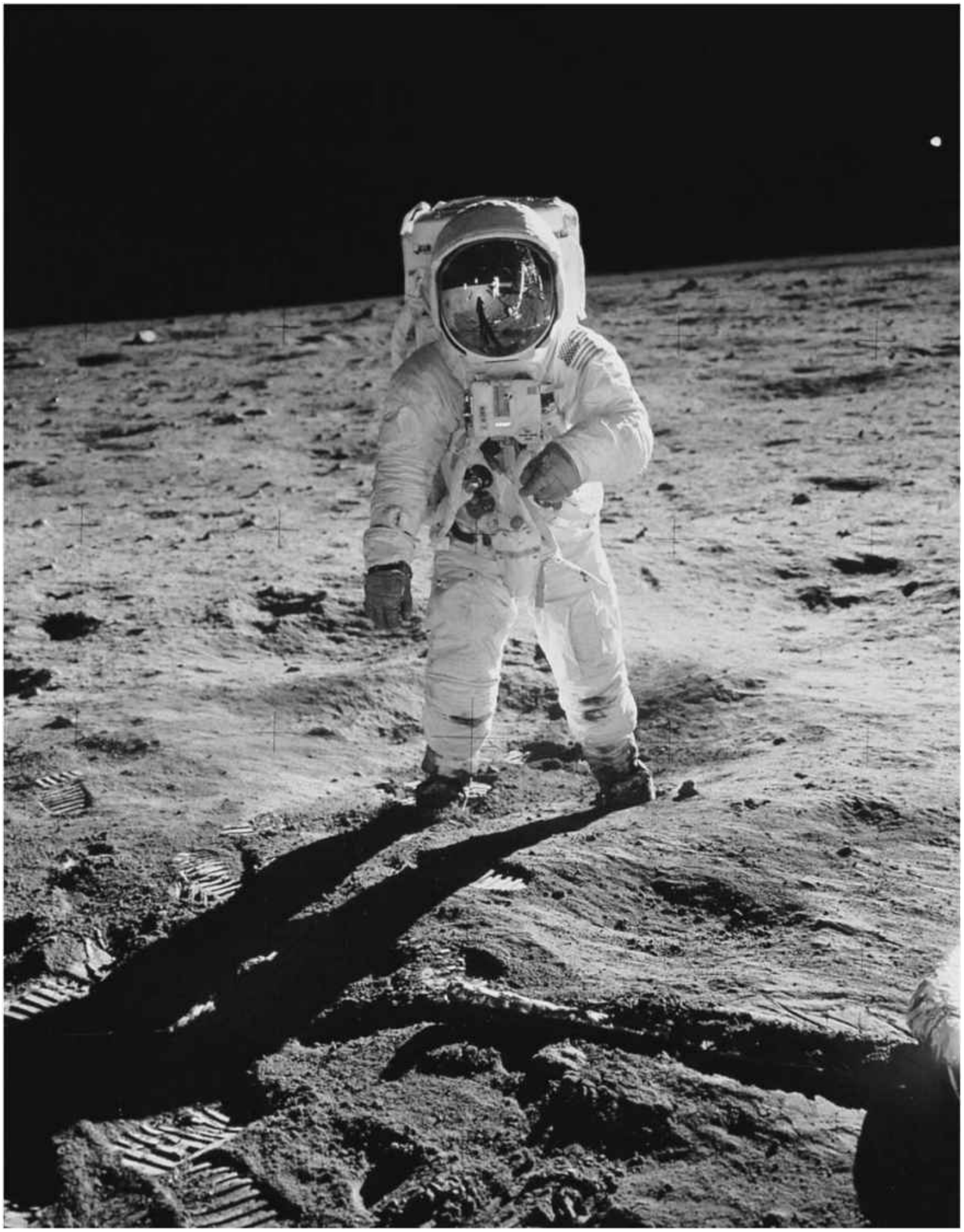
The Sun’ s light has over the course of millions of years enabled the formation of the current biosphere and the geology that provides other sources of energy. Neither coal nor oil could have been made without energy from the Sun. And our use of these resources is changing the way in which light from the Sun afiects our planet. Some of the invisible light from the Sun—ultraviolet light—is still absorbed by the Earth and its atmosphere. But the other invisible part—infrared light—is reffected back by atmospheric gases. By the same mechanism, infrared radiation is trapped on the planet, contributing to increased planetary surface temperature.

Light enables communication

Pictures have been part of human culture since the beginning of the species. The impact of images on how we conceive of the world, and how we make sense of our place in it, is immeasurable. Optical technology has contributed to this in ways that are utterly transformative. For instance the ability to capture images easily and rapidly, by means of film-based and digital photography, allows us to record places, people, and things as reports that can be widely distributed (nowadays by an optically enabled Internet) and which have lasting impact: images of leaders and workers, awe-inspiring scenes of the natural world, and horrifying

scenes of war. These can bind or fragment people in unexpected ways: calling populations to action, inciting acts of compassion, and giving deeper insight into shared experience. Recall the astonishing sense of wonder the sight of man’s first steps on the Moon (see Figure 2) invoked. The ability to capture moving images adds a completely new dimension by enabling new narrative and documentary capabilities. Can you imagine life without television or movies, without video?

Nowadays, the generation and transmission of images is so prevalent we hardly think of it. We use self-luminous displays every day: televisions, computers, tablets, even smartphones. All of these bring information to you and receive it from you using light as a medium. Almost all long-distance telecommunications travel on light beams, guided along thin strands of glass called optical fibres. This is the basis of fibre optical broadband services that link our homes to the Internet. Even inside computers and televisions light plays a role. For instance, the music, video, or images that are locked in a CD or DVD are accessed using light. A tiny moving head based on a miniature laser ‘reads’ the disc and converts the information coded on it into electrical signals that can then be sent to the display screen. All our surfing, downloading, and emailing activity now requires such immense information capacity that light is the only feasible medium for conveying it.



2. Neil Armstrong’s photograph of Edwin Aldrin walking on the Moon.

Transportation in the modern world uses light as the means by which we signal and regulate our movements. From streetlights in towns, to landing lights on aircraft, light is an essential part of navigation. And it even plays a role in maintaining our vehicles. For instance, lasers are used to align the wheels of a car, and light-enabled distribution of ignition power to drive an internal combustion engine is not uncommon.

In many, many ways, light carries the energy and information that makes modern life possible.

Optics

The field of enquiry that constitutes the study of light is called optics. Optics is among the oldest of the sciences, and its historical development forms one of the most important paths in the emergence of modern science. Ideas arising in optics have stimulated new ways of thinking for very disparate fields, such as the mechanics of motion of atoms and molecules. And technologies enabled by deeper understanding of light have been central in unlocking other secrets of the natural world. Galileo’s telescope designs, for instance, were critical in his observation of the moons of Jupiter, which was a vital step in moving towards a view of the solar system in which planets moved around the Sun. This, in turn, was important in developing the laws of gravity that govern planetary motion.

The origins of optics lie in the work of the Greek philosophers of the 4th century bce, and the field has continued to fiourish for the past two millennia. It is perhaps surprising that we can still discover new things about light after such a long period of attention by many clever people. Yet optics remains at the forefront of current science: more than ten Nobel Prizes in the past twenty years have been awarded for research in which light has played a central role, from controlling and measuring the motion of atoms and molecules, at unimaginably low temperatures and on timescales of breathtaking brevity, to improving the precision of clocks a thousandfold, to enabling us to look inside living cells and watch what happens as they change.

What is light?

A good place to start is with some of the things commonly associated with light: brightness, intensity, colour, and warmth.

These are all tangible properties that suggest light is a physical entity. But what, exactly, is it?

We may consider what it means for light to be bright by taking a particular source of light—a household light bulb. These come in various sizes, but all have powers of several tens of Watts (the unit in which power is measured, labelled W, and signifying the energy consumed per second of operation). A 50 W light bulb gives suffcient light by which to see things inside a house. Car headlights are typically of

slightly greater power, approximately between 60 W and 100 W. The fioodlights at a football ground are of much larger power, up to several thousand Watts. I will discuss later exactly how these different sources generate light, but these powers will give a sense of the brightness of the corresponding lighting. Of course, one of the brightest sources is the Sun. It has a massive power output, more than 10^{25} W (1 with 25 zeroes after it), which makes it impossible to look at directly even though it is a very great distance away.

This brings us to the next concept associated with how bright a light is. Light of the kind discussed above looks dimmer the more distant it is. So power alone is not the only criterion determining brightness. It is related in some way to the fraction of the power that we can receive from the light source. For example, a laser pointer typically has much, much lower power output than a light bulb, often only a few thousandths of a Watt (10^{-2} W or 10 mW). Yet it appears very bright when it is pointed at a screen.

What is important here is the intensity of the light generated by the source—the power per unit area of the receiver. (This is more properly the irradiance, but intensity is perhaps a more familiar term.) The intensity of a light source is related to the ability to concentrate the light. A laser pointer appears to be very bright because its light beam is concentrated on a small spot on the screen, whereas the Sun’s light is difiused over a very wide area.

Therefore, although the Sun has a very great power output, the light it produces is not as intense as that of a laser pointer.

The underlying property that describes our ability to concentrate light is called the source coherence. This is related to the source’s propensity to send light in a particular direction. For instance, both the Sun and a light bulb radiate light into all directions—the Sun can be seen from every place on Earth, and a light bulb from wherever you stand in a room. But a laser pointer only puts out light in a single direction—that in which you point it. So you cannot see the laser beam unless you look at the surface on which it is incident. Because of the property that its beam has a well-defined direction, the laser pointer is said to be a coherent light source, whereas the light bulb is an incoherent light source.

Another defining characteristic of light, perhaps its most evident property, is colour. The rainbow embodies the fundamental idea of a spectrum of colours—a palette emerging from the conjunction of rain and sunlight—from blue at one end to red at the other. A central strand in the development of theories of light has been the development of models of colour vision. Colour is intimately tied to perception, as well as to physics. An illustration of this is found in the experiments on the nature of colours undertaken by Sir Isaac Newton (Figure 3), a dominant figure of early 18th-century science, and whose book Opticks defined the view of light for

two centuries, and by Johann Wolfgang von Goethe (Figure 3), a dominant figure of late 18th-century literature, who incorporated scientific ideas in his writings, but nonetheless believed Newton to be profoundly wrong about the nature of light.

The first part of Newton's famous experiment (similar to ones undertaken by Descartes and others previously to him) was to allow a small beam of light from the Sun, defined by a hole in a dark screen, to pass through a prism, and thence to fall upon a screen. The familiar rainbow colours emerge. Goethe was fascinated by this effect, and borrowed some prisms from a local aristocrat with which to experiment for himself. He concluded quickly that Newton's experiment was wrong—in the sense that his claim of the universality of colour dissection from white light was not true. Goethe had himself discovered a very different set of colours.



3. Isaac Newton (left), Johann Wolfgang von Goethe (middle), and Rosalind Franklin (right).

The experiment Goethe undertook was to look through the prism at the mullion of a window. That is, he looked at a dark line against a bright background, the very opposite of what Newton had done. What he saw was a very different spectrum from Newton. Not the red, green, and blue of the Newtonian spectrum but rather a new palette of cyan, magenta, and yellow; the so-called complementary colours of the Newtonian spectrum. Combining Newton's colours produces a white image—combining Goethe's produces a black one.

Goethe espoused the idea that colours are the things perceived; Newton defined them as intrinsic properties of light. They were both right. Nowadays we are content to separate the physical attribute of colour from its physiological effects, the sensation of colour. We each react differently to colours—indeed, coloured light can even be used as a therapy. From an artistic point of view, the interpretation that our consciousness realizes from the sensations associated with light of a particular colour is a critical matter—perception is vitally important. Yet from a physical point of view there is an underlying property that we can assign to the label 'colour' unambiguously: its frequency—at least until we get into the realm of quantum light.

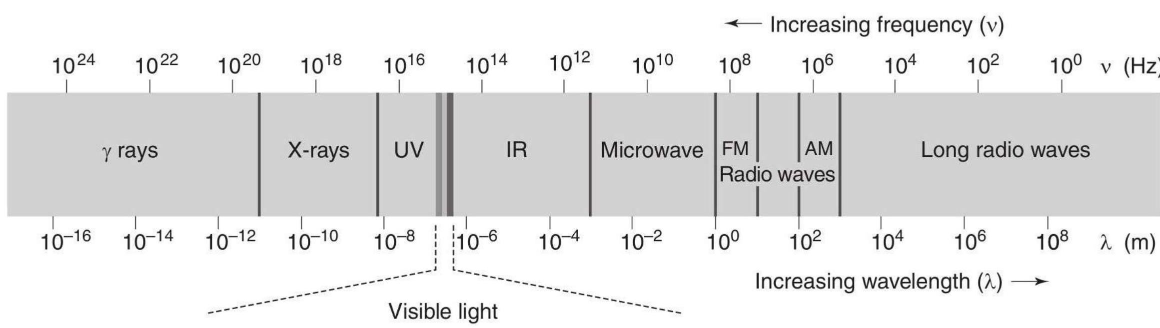
The reach of light extends beyond the visible spectrum, at the blue end into the invisible realm of the ultraviolet and the extreme ultraviolet to X-rays and γ-rays. At the other lie the infrared, microwaves, radio waves, and

eventually T-rays (Figure 4). To see these, we need different instruments than our eyes alone. Nonetheless, we know that light of these colours exists. For instance, the warmth of the Sun is due to infrared light that is absorbed by our skin. At lower frequencies microwaves are used for cellular phone communications as well as for cooking, by heating the water in the food being prepared. The invisible colours at shorter wavelengths are also familiar. Sunburn is caused by ultraviolet light, while X-rays are used routinely in medical imaging.

X-rays are also used in many non-medical applications. For example, the patterns made when X-rays scatter from a regular arrangement of atoms in a molecule or solid enable that arrangement to be determined, even though the atoms are spaced very, very close together, at distances more than 10,000 times smaller than a human hair. X-ray diffraction images can reveal the very structure of the molecule or solid, with profound implications. Perhaps the most famous example is the structure of DNA molecules identified by James Watson and Francis Crick more than half a century ago, based on X-ray images taken by Rosalind Franklin (Figure 3) and Maurice Wilkins. Knowing how molecules replicate revolutionized biomedicine.

These applications are indicative of the importance of light—in its broadest sense—in making possible our modern world, and impact our ability to enjoy it to its full. They rest on the fundamental work of a number of scientists in the

19th century—Michael Faraday, Hans Christian Oersted, André-Marie Ampère, Charles Augustin de Coulomb, Alessandro Volta, Georg Ohm, James Clerk Maxwell, and Heinrich Hertz. That there exists a connection between visible light and other apparently disjoint things like microwaves and X-rays is remarkable, and it was the triumph of scientific enquiry by these men and others who made and identified these connections.



4. The spectrum of electromagnetic waves.

The gamut of colours, or spectrum, provides a tool for art and science. Whereas a painter or artist explores ways in which colours themselves are juxtaposed or combined, a spectroscopist explores ways in which matter responds to different colours. For example, in the early 19th century Joseph von Fraunhofer determined some of the types of atoms that are present in the Sun by looking carefully at the particular colours of light that the Sun emits. He noticed that characteristic colours were missing from the Sun's spectrum and noted that these colours were 'fingerprints' for particular atoms. The study of spectra is the domain of

spectroscopy, which uses light to identify different atoms or molecules. It is a vital research activity today, with impact in many areas from health monitoring to remote sensing of the atmosphere for pollutants.

Apart from these familiar properties of light, there is one further property I want to point out. It, too, is something that we all know about from our everyday lives, though perhaps in a less explicit way than we experience the other properties of light. It is polarization.

If you have watched a 3D movie, then you have seen this property being exploited. Watching such a movie requires you to wear special spectacles with cardboard or plastic frames that have pieces of plastic for the 'lenses'. If you take two pairs of these spectacles and slide the left lens of one over the right lens of the other, then look through them at a light bulb, you will see a very, very dim image of the bulb. Alternatively if you rotate one of the spectacles through 90 degrees with respect to the other and put either the two left lenses or the two right lenses over one another you will see something very similar—almost no transmission of light through the pair of glasses.

This can be explained by assigning a property of 'orientation' to the light. Most common light sources emit light without a preferred orientation. When you look at a light through these spectacles it will appear dimmer, indicating that they have selected a particular orientation.

The left lens allows one orientation to pass, the right lens an orientation that is at right angles to the first. That’s why when you put a second lens oriented at right angles to the first, nothing will be transmitted through it: the light passing through the first lens has the ‘wrong’ orientation to pass through the second. This characteristic of orientation is called polarization. It took a great deal of careful exploration to come up with and to understand the idea of polarization, but it is an important feature for technologies based on light, and indeed for understanding what light actually is.

These physical characteristics of intensity, colour, and polarization are what enables light to be used to discern, to measure, and to control properties of material substances, and therefore to form the basis of a host of tools for studying and manipulating matter and even small objects. In the examples above, light almost always plays the role of an information carrier. Whether it is conveying an image or a spectrum or a phone conversation, it acts as a messenger. But light can also be used in an active way. For instance, the heating properties of light can be used to cut metal and other materials very precisely. Thick plates of metal, up to a centimetre or so, can be machined using high-power lasers more rapidly, with a higher quality finish and less waste, than using a saw. And light is used in medicine in various ways, from correcting vision through laser surgery to activating drugs for anti-cancer therapies.

Light enables us to view the natural world at every time scale and distance scale imaginable; from the earliest moments of the universe to the unimaginably fast motion of electrons in atoms and molecules; from the large-scale arrangement of clusters of galaxies across the universe to the atomic arrangement of carbon atoms in graphene. It provides us with insight into the very foundations of the natural world, from the weirdness of quantum physics to the structure of DNA molecules.

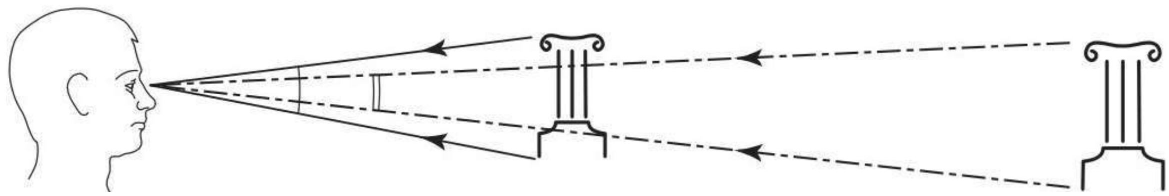
The story of optics is one in which new discoveries about light have enabled new technologies that have, in turn, given rise to new discoveries across many fields of science. At each stage, from the invention of eyeglasses to the most precise atomic clocks, to modern technologies for imaging, measuring, and communications, light has given new applications that have revolutionized how we live. This cycle of discovery and innovation makes the study of light a vibrant discipline, even as it is among the most venerable. How we came to a modern understanding of what light is, and therefore how we could use it, both for new understanding of the world and for new capabilities that change the world, is the tale that forms the rest of the book.

Chapter 2 Light rays

Taking a selfie requires that the phone camera is pointed towards you. It' s obvious that this is necessary if you are to be in the picture. But this simple fact indicates something about the nature of light: to see an image of an object, there has to be a straight line between the object (in this case you) and the camera lens. This is usually called the 'line of sight' . Thus light is something that propagates in a straight line from the object to the viewer.

Indeed, this is what we might expect from our knowledge of certain types of light source. Eye-catching visual effects at concerts are generated using lasers, illuminating the stage and the performers with coloured light beams. Laser pointers are commonly used at talks or lectures to emphasize images or words on a screen. The beam produced by these coherent light sources is highly focused, hardly diverging at all even across a big hall. It goes in a straight line—you point the device in the direction where you want the light to go, and it does so.

Because sunlight does not obviously exhibit this characteristic, it required some thinking to determine that the propagation of light in a straight line was exactly what was needed to understand why distant objects appeared to be smaller than nearer ones, even when it was known that they were, in fact, exactly the same size physically.



5. Euclid’ s construction of rays showing why objects of the same size look smaller when they are further away.

The insight that the concept of straight-line propagation could explain this effect is attributed to Euclid, working in Greece in around 300 bce. His idea, from one of the earliest books on optics, is illustrated in Figure 5. Imagine two lines—let’ s call them rays—one connecting the top of the object (a pillar in this case) to the observer’ s eye, and one connecting the bottom of the object to the eye. The angle between these is related to the apparent size of the image of this object that we perceive. A more distant pillar, the same physical size as the first, produces two rays whose angle of intersection at the observer is smaller—hence the pillar appears smaller. This is what we call perspective in the image.

What is the stuff that traverses the rays? Euclid (building on earlier ideas) thought that it was particles sent out from the eye itself (from an imaginary internal fire) that illuminate the object and are reffected back to the observer. But this suggests that we would be able to see things whether or not it was dark outside. Nonetheless, the idea of stuff moving along a trajectory between object and observer remained a powerful concept.

It was modified by Al-Hazen in the 11th century to a form that we now use routinely: objects are illuminated by rays from the Sun (the external fire) and are scattered towards the observer. There are several stories about how he came to this idea—including doing an experiment when he looked directly at the Sun, and determined that the painful sensation he experienced would be there all the time if the 'internal fire' were burning all the time. Thus, he argued, the source of the light necessary to generate the image was external.

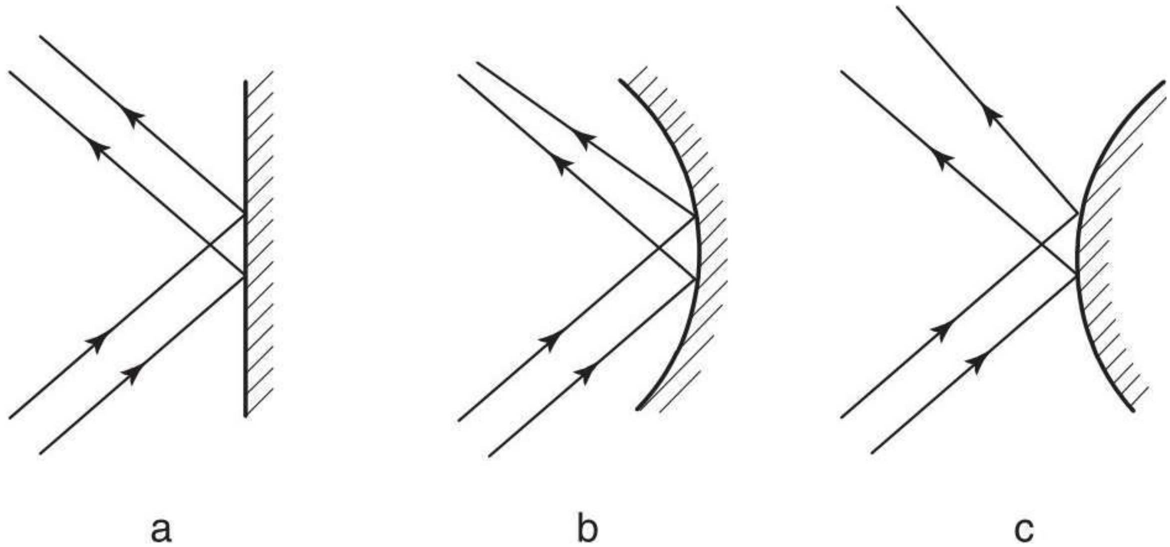
Let us, for the sake of this argument, assume that what moves along these rays are particles of light: call them photons. The brightness of the beam is related to the number of photons traversing the ray in one second. In order to understand how an image of an object is formed, we'll need to consider what happens when one of these photons is reffected off a mirrored surface, as well as what happens at a lens. This will lead to the 'laws of optics' that are used for designing very complex optical instruments such as surgical microscopes, and catheters for 'keyhole' surgery, as well as massive optical telescopes placed in orbit above the Earth for observing distant galaxies. The impact of these instruments on our life and our understanding of the world is immense.

What sort of properties do these particles of light possess? The usual sorts of attributes assigned to a particle are its position, its direction of travel, and its speed. For

now, assume that it moves at the ‘speed of light’, without going into detail about what that actually is. The position of a photon might then specify the starting position of the ‘ray’ and the direction of the photon’s motion would be the direction of the ray. The photon heads off from its starting point in this direction at the speed of light until it encounters the surface of the object.

Reflection

When it hits the object, the light is reffected. What happens is that the photon ‘bounces’ off the surface, changing its direction, but not its position on the surface. The situation is shown in Figure 6. The manner in which the direction is changed is specified by the ‘law of reffection’, discovered by Hero of Alexandria in the 1st century. It states that the angle of incidence (that is, the angle between the incoming ray direction and the direction perpendicular to the surface at the point of incidence) is equal to the angle of reffection (that is, the angle between the outgoing ray direction and the direction perpendicular to the surface at the point of incidence). There are some surprising consequences of this conceptually simple and yet extremely powerful law.



6. Rays of light reflecting off a. a faat mirror and b. and c. a curved mirror.

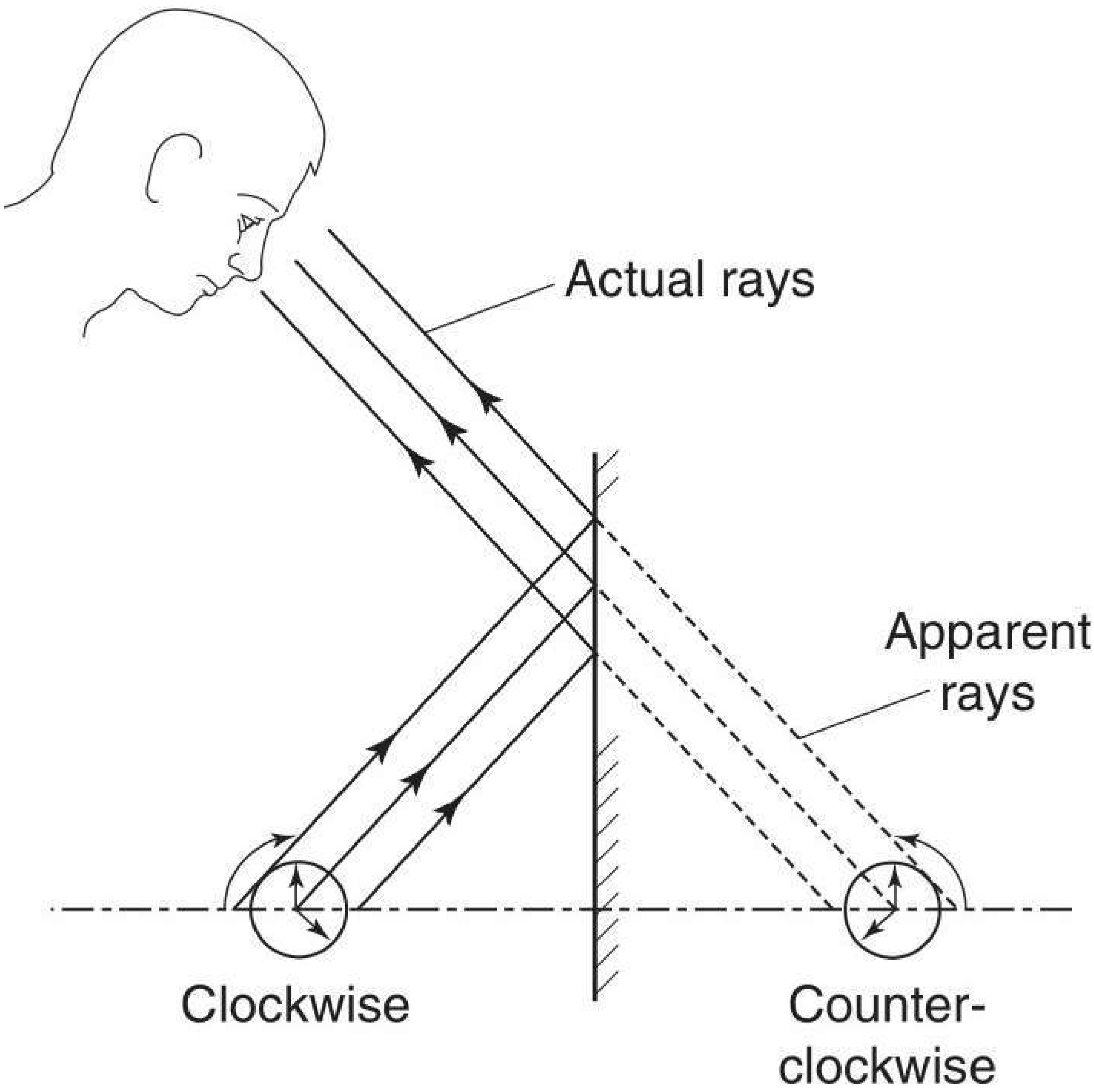
The apparent left- right inversion associated with reffections follows from this idea. If you hold up your watch to a mirror and look at its reflection, you' ll see that its second hand is moving in an anticlockwise direction, and that when you move your right hand, your doppelgänger moves his or her left hand. This change of 'handedness' is the hallmark of the image through the looking glass—the world reffected in a mirror is in that sense really topsy-turvy.

This can be explained entirely by Hero' s law of reflection. Figure 7 shows how a mirror generates a reflection with opposite handedness. The clockwise-pointing arrow is the object. Rays from each point on the arrow reflect off the mirror and are rearranged so that the arrow seen in the mirror is pointing anticlockwise. You can use the same construction to show that a left-pointing arrow is seen

in the mirror as a right-pointing arrow and vice versa, but an up-pointing arrow remains pointing up, and a down-pointing arrow points down in the reflection.

Imaging using reflections

If, instead of looking at your opposite-handed self in a flat bathroom mirror, you see your reflection in a polished spoon, then you see a distorted form of your image: magnified features on curved backgrounds. The concave front surface of the spoon magnifies the object and the convex back surface demagnifies it.

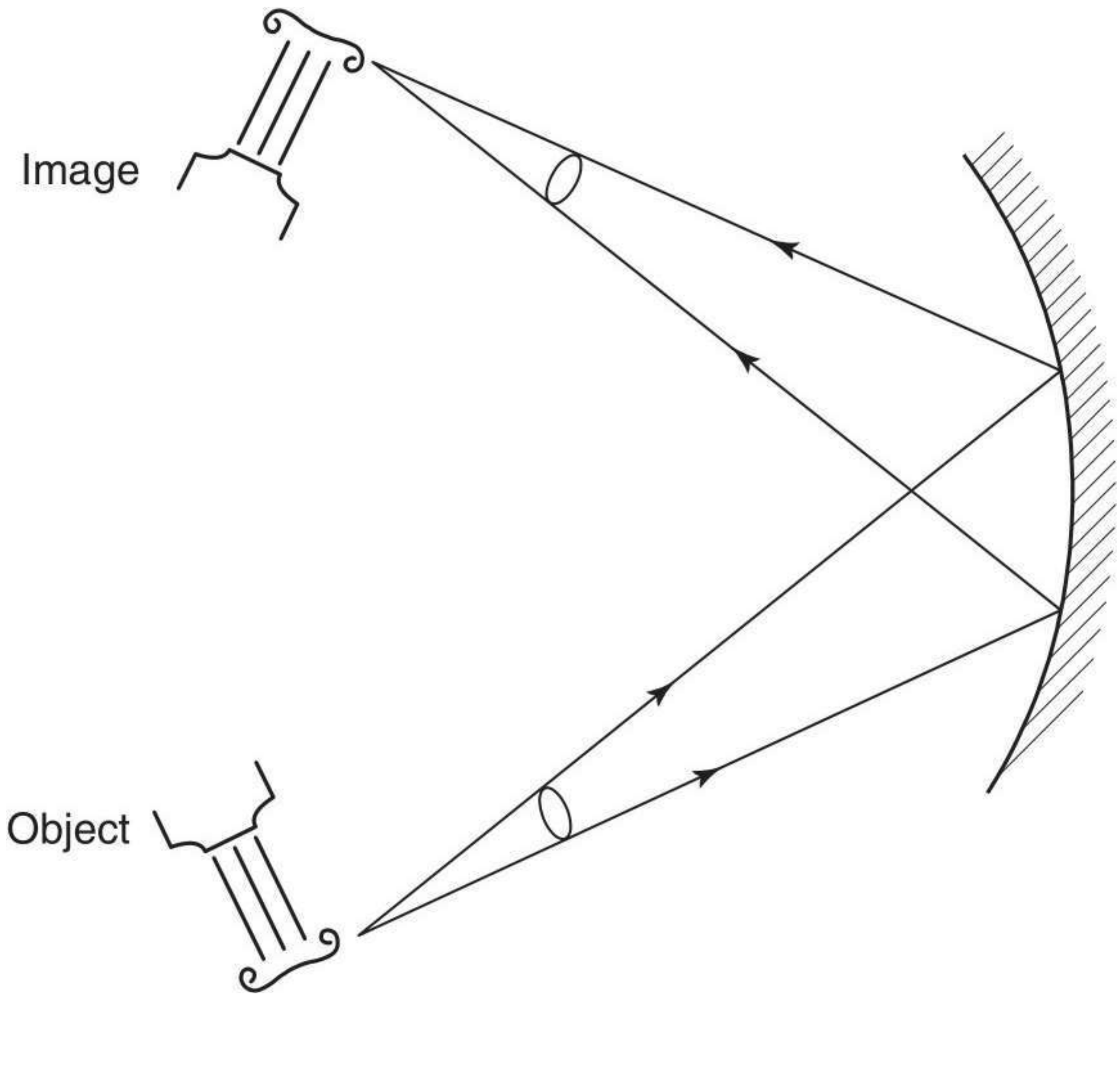


7. The change of handedness of objects reflected in a mirror.

Why is this? Since the formation of images is perhaps the most important application of optical instruments—ranging from contact lenses for vision correction to space telescopes for scientific discovery—it’s worth understanding exactly how this happens.

Up to now, I’ve considered only a single ray of light from one point on the object. In fact, rays usually scatter

in all directions from each object point. Consider a 'bundle' of such rays all coming from a single point on the object, forming a cone around the original ray. This cone of rays diverges as it moves away from the object, as shown in Figure 8. The rays hit the curved mirror surface at different points and therefore also at different angles of incidence. Thus they reflect in different directions, each one still satisfying Hero's law of reflection.



8. Image formation by a ray bundle using a curved mirror. The rays from one point on the object meet at one point in the image.

In fact they now form a converging cone, and eventually all meet at a single point. This is the ‘image’ point of the original object point.

An image as we normally consider it is made up of multiple such image points arising from different object points. The size of the image is determined by the distance of the object from the mirror and the focusing power of the mirror, specified by the radius of curvature of surface (e.g. a more concave mirror has a smaller radius of surface curvature). The image can be bigger than the object when the object is closer to the mirror than the image. The ratio of image to object size is called the magnification.

The image-magnifying feature of a curved mirror was used by Newton to design a telescope, shown in Figure 9. His design has a remarkable property: it forms images of distant objects that are the same size for every colour (said to be free from ‘chromatic aberration’). Newton cleverly used the property that the angle of a reflected ray for a fixed angle of the incident ray is the same no matter what the colour of the light. The image of each colour is therefore formed in the same place—all the colours register perfectly, guaranteed by physics.

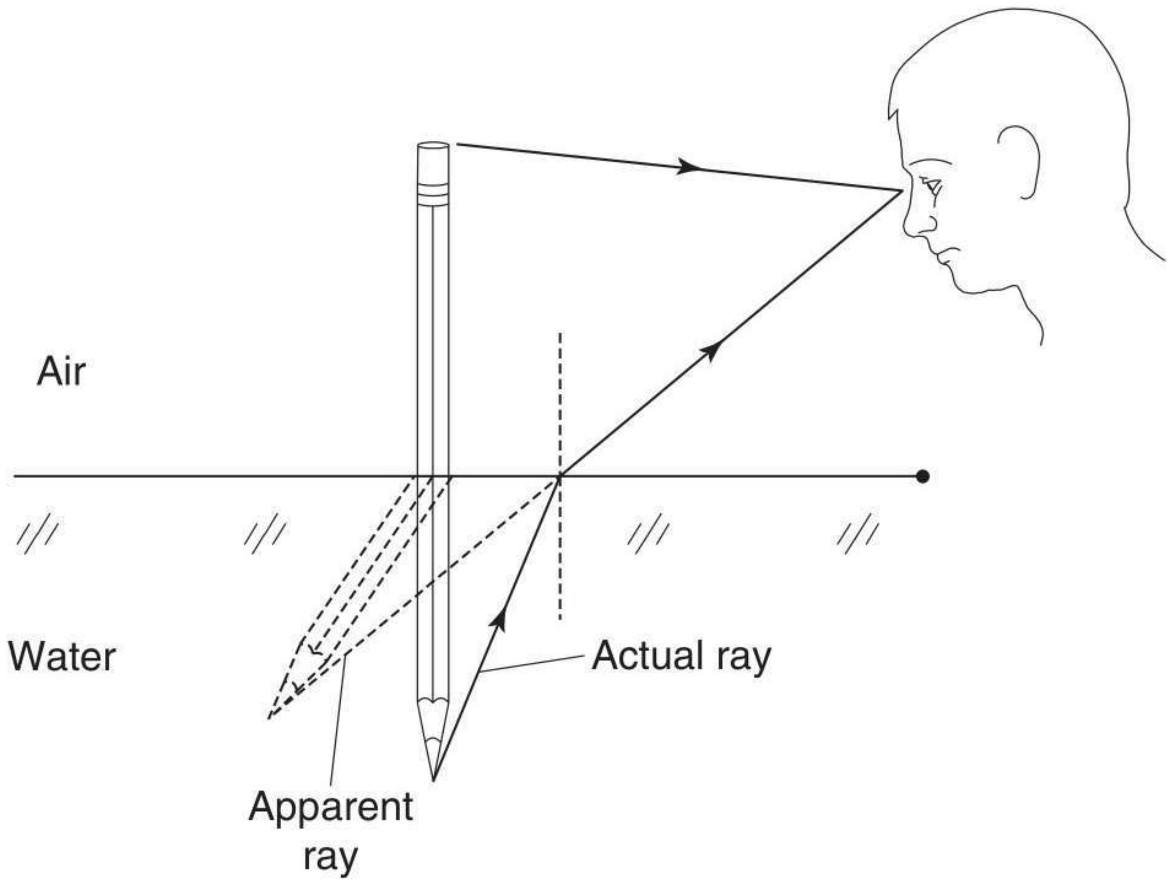


9. Newton’s reflecting telescope enabled images without chromatic aberration.

Refraction

Newton invented this instrument because the telescopes used by contemporary pioneers such as Galileo Galilei and Johannes Kepler suffered seriously from chromatic aberration. The images formed by their telescopes always had a blurred coloured halo around the edge of the object. The reason for this is that they were designed using a different property of light rays—refraction: the phenomenon that light rays bend when they go from one transparent medium to another.

It is the refraction of light that produces the 'kink' observed in a pencil partially immersed in a bowl of water. This is described by the law of refraction, commonly known as Snell's law, after Dutchman Willebrord Snell, an early 17th-century proponent. This law says that the angle the exiting ray makes with a line perpendicular to the surface is related to the angle of the incoming ray by the properties of the two media that form the interface—in our example the surface of the water. This is illustrated in Figure 10.



10. Refraction of a ray at the interface between air and water.

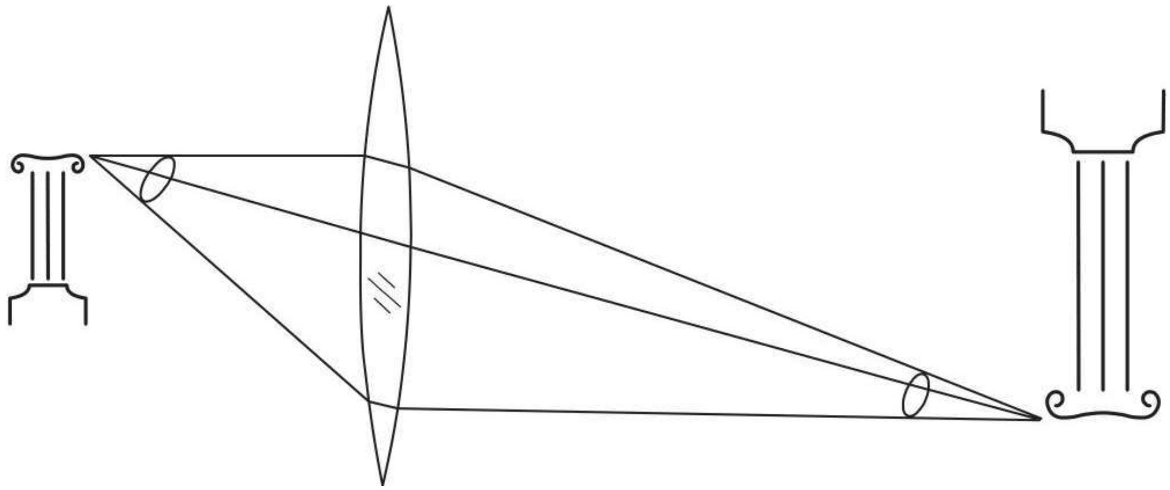
The particular property of the media that is relevant here is the ‘refractive index’. The refractive index can be thought of as being a measure of the optical ‘stiffness’ of the medium as experienced by a light ray. So light travels more slowly in a medium with a larger refractive index because the molecules of the medium are slightly more resistant to having their atoms and electrons moved by the light. It is like running in a pool of water. If the depth is very small, your legs can move easily and you can run fast. If the water is up to your knees, it’s harder, because you have to work against the resistance of the water.

In fact, the law of refraction can be derived from this analogy. Pierre de Fermat showed that when light goes from a point in one medium to a point in another medium, it seeks a ray that traverses a shorter time in the medium with high refractive index and a longer time in the medium with a lower refractive index. This requires the light ray to bend at this interface between the two media, and Fermat's principle turns out to be entirely the same as Snell's law.

Imaging using lenses

Now, just as a curved reflecting surface can form an image of an object, so can a curved transparent surface. How this happens is shown in Figure 11. A bundle of rays from a point on the object are brought to focus at the image. Notice the shape that does this—it has the same cross section as a lentil. This is the origin of the word 'lens'.

Lenses are ubiquitous in image-forming devices, from your eyes to mobile phone cameras to surgical microscopes. Imaging instruments have two components: the lens itself, and a light detector, which converts the light into, typically, an electrical signal. In the case of your eyes, this is the retina, whereas for the mobile phone it is an array of minute pieces of silicon that form solid-state light sensors.



11. Image formation by a ray bundle using a lens.

The lenses in each of these devices are different, of course, but the basic principle is the same for each. In every case the location of the lens with respect to the detector is a key design parameter, as is the focal length of the lens which quantifies its 'ray-bending' power. The focal length is set by the curvature of the surfaces of the lens and its thickness. More strongly curved surfaces and thicker materials are used to make lenses with short focal lengths, and these are used usually in instruments where a high magnification is needed, such as a microscope.

Because the refractive index of the lens material usually depends on the colour of light, rays of different colours are bent by different amounts at the surface, leading to a focus for each colour occurring in a different position. This gives an image that has 'haloes' of different colour around it. For example, only one colour may be properly in focus at a particular detector plane; the others will be out of focus,

and form the halo. Whether this chromatic aberration is an important effect or not depends on the particular application.

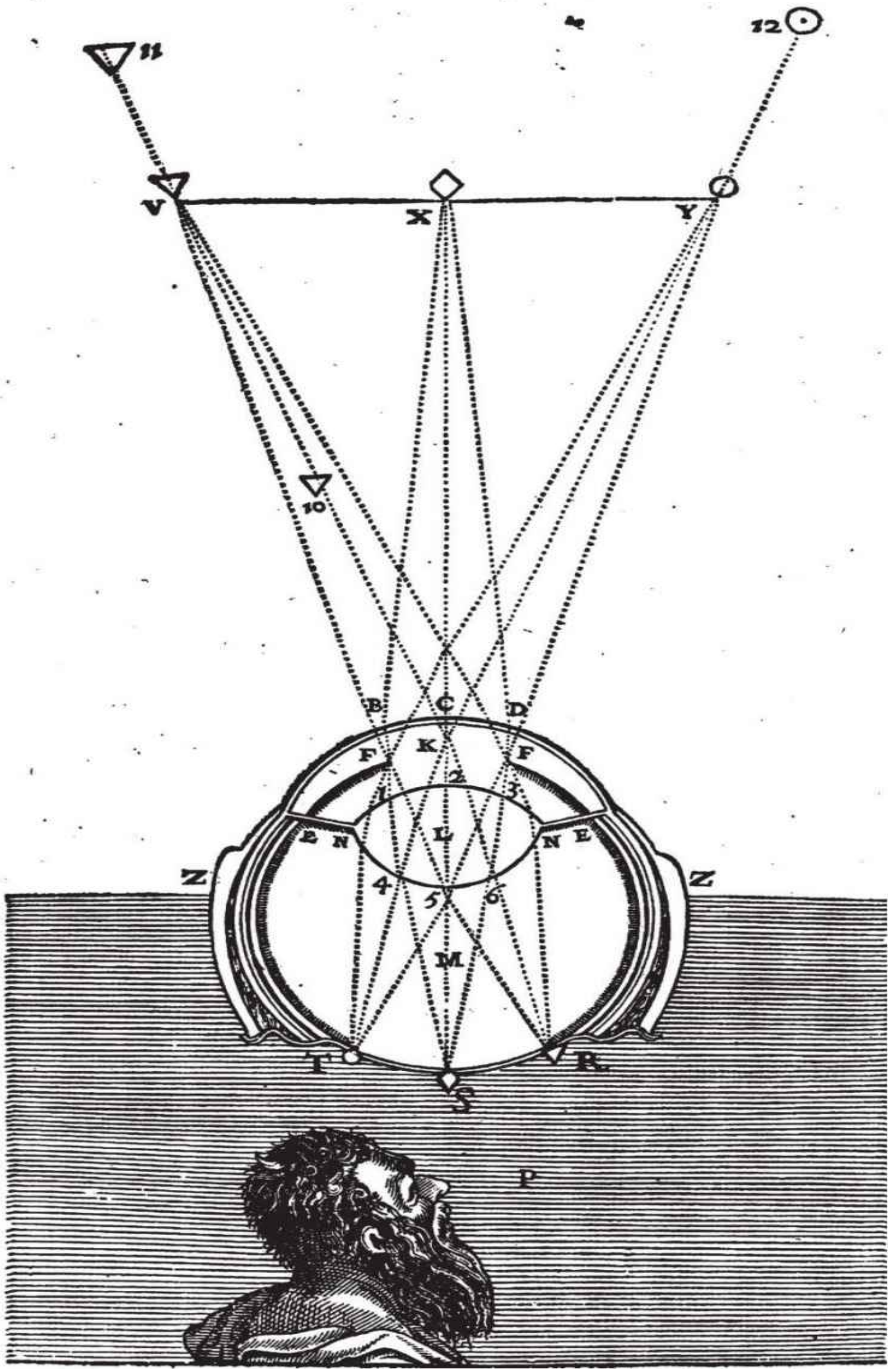
One of the most familiar, and indeed most important image-forming instruments of this type is the eye. It consists of a front-refracting surface—the cornea—and an adjustable lens, which changes shape as you focus on things at different distances. These elements form images on the retina at the back of the eye.

Historically, the formation of images by the eye was of great interest, since an experiment by Descartes (see Figure 12) showed that the image of an upright object was upside down. Of course, we don't perceive the object in this way, so it was clear that the brain undertakes some remarkable processing between the raw retinal signals and the perception of the external world.

Optical instruments

As many of us are all too aware, the ability of the eye to form high-quality images (sharp, undistorted, and in colour) can degrade as we age. One of the earliest applications of optical instruments was developed as an aid to sight under such circumstances. Eyeglasses were perhaps the first optical technology, purportedly invented by Roger Bacon—the 'mad friar' of Oxford—in the 13th century.

The corrective elements are often simple lenses, placed either in a frame at some distance (a few millimetres) from the cornea (the front surface of the eyeball), or ‘contact lenses’ placed, as the name suggests, in contact with the cornea. In both cases the imaging system is compound, that is, it consists of several elements—the external lens, the cornea and the internal ocular lens. This functional form provides the necessary degrees of freedom to correct most kinds of vision by enabling the external lens to compensate for the imperfections of the internal lens. This can also be done by directly altering the shape of the front surface of the eye by laser surgery. One approach, laser-assisted in situ keratomileusis (LASIK), uses the laser to ablate part of the surface of the cornea. This changes its curvature, thereby altering its focusing power, and thus the image-forming capabilities of the eye.


11
12
V
X
Y
10
B
C
D
F
K
F
1
2
3
E N
L
N E
Z
Z
4
5
6
M
T
R
S
P

12. Descartes' experiment to show the image formed by an eye is upside down.

Many other image-forming instruments work on very similar principles to the eye. The camera of a mobile phone, for instance, has a lens near the surface of the phone, and a silicon-based photodetector array inside the device. The mobile phone lens is often very small, and yet must provide images of suffcient quality that they are intelligible, and thus make sense when they are posted on Facebook. This requires that both the detector array and the imaging system are adequate to the task of producing a high-quality image. The quality of the image depends on two things—the size and the number of detectors in the array, and the ability of the optical system to create a sharp, undistorted image with all colours properly registered. That is, an image free from 'aberrations'.

The specification of the number of 'pixels' of the detector array is often used as a proxy for the quality of the image. A 24 megapixel camera (one in which the detector array contains 24 million sensors) is often considered better than an 8 megapixel one. A pixel can be thought of as the size of the image of a point object. If it is possible to resolve only a small number of points—because there are only a few elements in the detector—then it is hard to tell much about the object. The more pixels, then, the better. But only

if the imaging system can produce a point image that is smaller than one detector element.

Limits to imaging

In the 19th century, the German scientist Ernst Abbe devised a simple rule for the minimum size of any image that was applicable to all imaging systems then known. Abbe's criterion says that the size (S) of the image of a point object is proportional to the wavelength (λ) of the light illuminating the object multiplied by the focal length (f) of the lens and divided by the lens diameter (D):

$$S=1.22\times \lambda f/D.$$

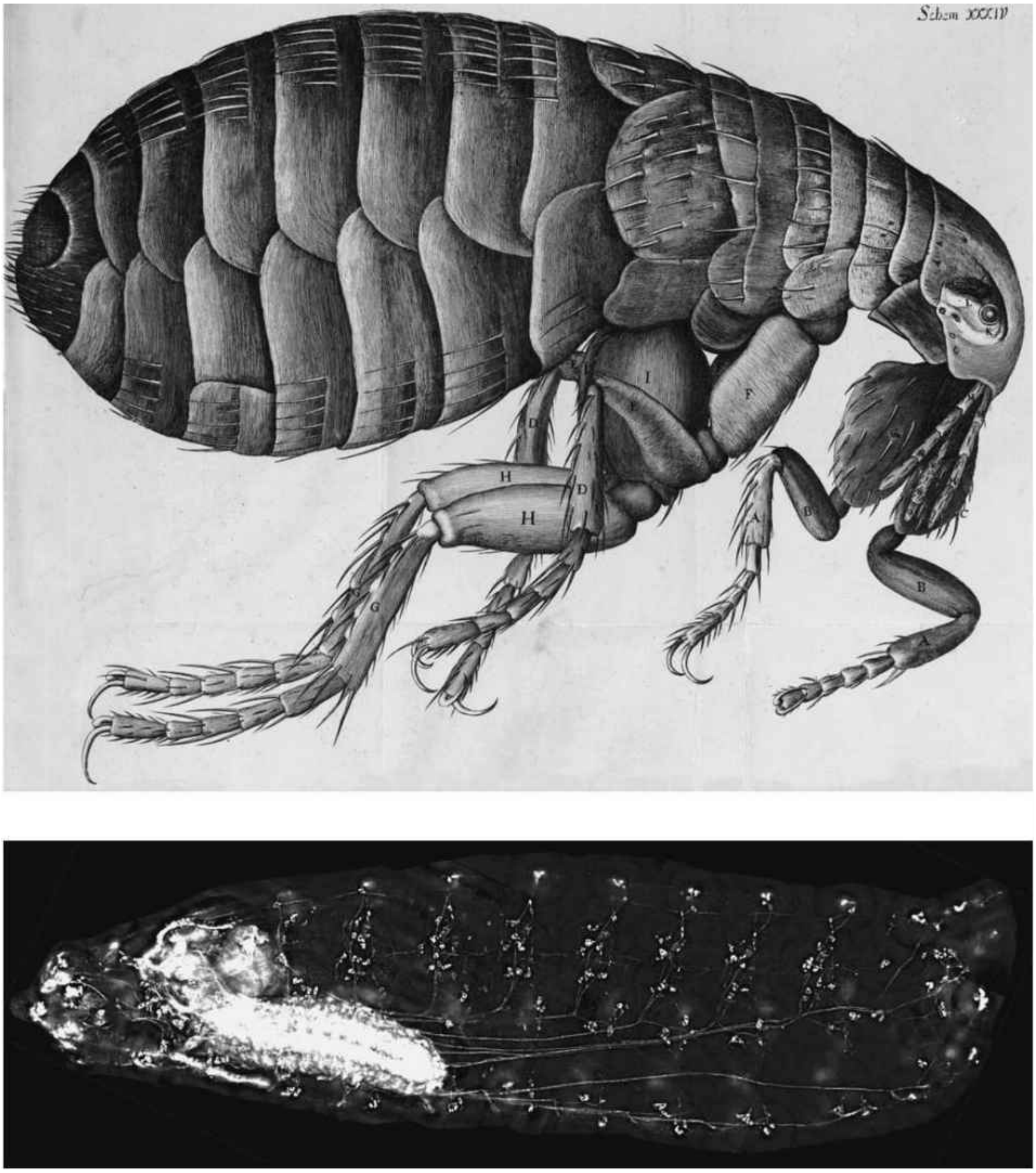
Thus, lenses with a big diameter and a short focal length will produce the tiniest images of point-like objects. It's easy to see that about the best you can do in any lens system you could actually make is an image size of approximately one wavelength. This is the fundamental limit to the pixel size for lenses used in most optical instruments, such as cameras and binoculars.

Designing and building imaging systems that can deliver high-quality images has been at the heart of many important applications of optics. Microscopes, for example, are used in applications from biological research to surgery. The earliest microscopes used very simple lenses—small, nearly spherical polished glass shapes that provided early experimenters like Robert Hooke in the 17th century with the

means to explore new features of the natural world, too small to be seen by the unaided eye. His drawing of a flea, shown in Figure 13, was a revelation of the power of technology to enable new discovery.

Modern research microscopes are much more sophisticated devices. They consist of a multi-element compound lens that can form images with pixels very close to the wavelength of the illumination light—just at the Abbe limit. Figure 13 also shows an example of what is possible using a modern imaging microscope. It is a composite image of the nervous system of a fruit-fly larva, about to hatch, made by viewing light-emitting proteins located in the cells.Abbe’s criterion applies to all optical systems in which the image brightness is proportional to the brightness of the object. These are called linear systems. But it’s possible to go beyond this limit by means of nonlinear systems, where the image brightness is proportional to the square or even more complicated function of the object brightness. A fuller explanation of these effects requires knowing a bit more about the wave model of light, which is the subject of Chapter 3.

Optical imaging systems of similar properties and complexity are used in another imaging application—the making of computer chips. Individual electronic circuit elements are extremely tiny.



13. Hooke’s diagram of a flea observed by means of an early microscope (top), and the nervous system of a fruit-fly larva (bottom) taken with a modern fluorescence microscope.

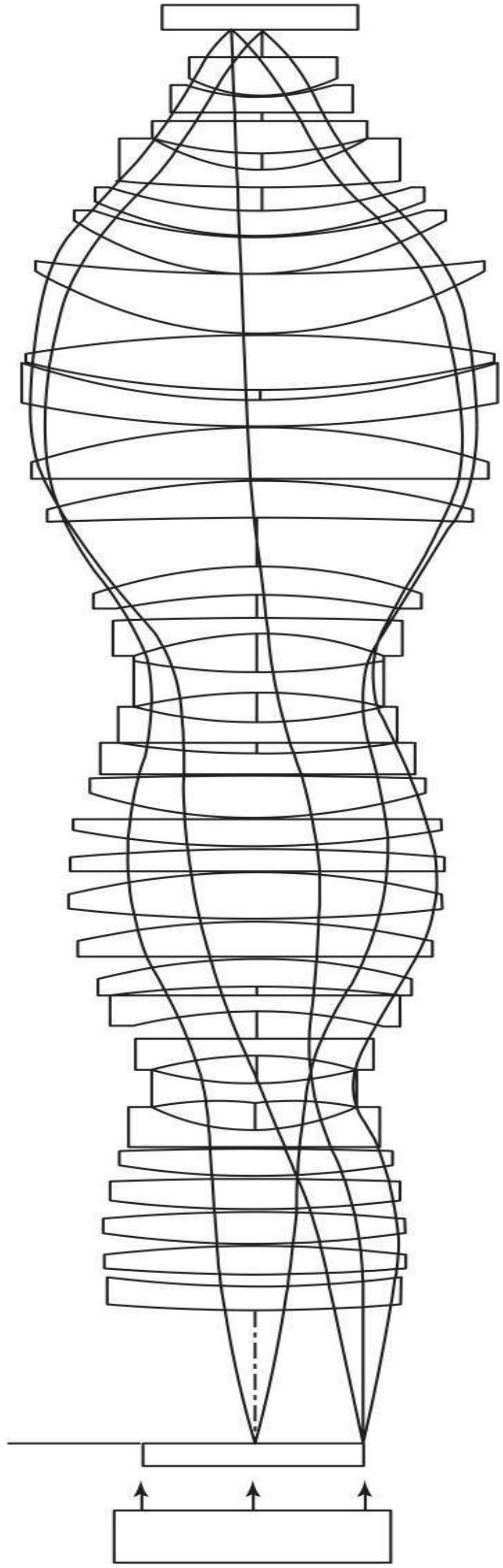
A wire connecting two transistors on a chip may have a diameter of only 250 nanometres (nm). (A nanometre is one

billionth of a metre, or 10^{-9} m. For comparison, a human hair is approximately 10,000 nm in diameter.) The complex array of devices and connections is laid out on a silicon wafer by means of a process called lithography.

Essentially, the chip layout is drawn at a large enough scale to be visible to the human designers, then a demagnified image is projected on to the chip. The image is etched into a surface coating on the wafer, and a series of chemical processes then maps the image into real devices. The imaging system must be able to provide extraordinary resolution in the image—with pixel sizes of the order of the line size of the device. Maintaining this resolution over an entire wafer is a real challenge, requiring many lens elements properly designed to reduce all aberrations to an absolute minimum. An example of such a lens is shown in cross section in Figure 14, showing the multiplicity of lens elements and ray paths.

At the other extreme, both ground- and space-based telescopes for astronomy are very large instruments with relatively simple optical imaging components, often consisting of just one curved reflecting surface and a simple 'eyepiece' to adjust the rays so as to make best use of the available detectors. The distinctive feature of these imaging systems is their size. The most distant stars are very, very faint. Hardly any of their light makes it to the Earth. It is therefore very important to collect as much of it as possible. This requires a very big lens or mirror—several

tens of metres or more in diameter. It is not practical to build lenses of this size, hence the ubiquity of mirrors in large telescopes. It is also necessary to look at distant stars for a long time in order to gather enough light to form an image. And this leads to another problem for ground-based telescopes: the atmosphere is not static. It changes in density with wind, temperature, and moisture. These fluctuations tend to make rays deviate from their course from star to telescope, causing the star to 'twinkle' as its light is deflected randomly onto and off the detector due to atmospheric turbulence.



14. Section of a lens used for photolithography of computer chips. It consists of more than twenty different lens elements, and produces images of 500 nm in size using light of less than half that wavelength.

One way to deal with this problem is to put the telescope outside the atmosphere, in space. The Hubble Space Telescope is an example. It produced spectacular images of distant stars, galaxies, and nebulae, showing extraordinary structures and movement in the far reaches of space. However, optical engineers have in the past two decades devised a clever way to deal with this problem for ground-based telescopes operating with visible light. What they do is to make the telescope mirrors in segments, the tilt of each segment being adjustable. Thus it is possible to 'steer' the rays hitting different parts of the telescope mirror so that they all hit the detector. If you can measure the deviation that a ray experiences as it traverses the atmosphere, then you can configure the mirror to compensate for that deviation. And that's what the engineers do. They measure how light from a guide star—an artificial light source in the upper atmosphere—is distorted and use that information to adapt the tilt of the mirror's segments. In this way images right at the Abbe limit can be produced. Space telescopes are still needed though, to probe the wavelength ranges, such as X-rays and UV, that are absorbed by the atmosphere, and several missions for new ones are

planned by the North American Space Agency (NASA) and the European Space Agency (ESA).

Metamaterials and super lenses

For many years, optical scientists have been fascinated with the question of what makes a good optical system. Is there a lens that can form the perfect image of an object? This question has intrigued many great physicists, from James Clerk Maxwell in England in the 19th century to Victor Veselago in the USSR in the 20th century. Veselago thought about materials that behave in strange ways—in which light bends in a way opposite to that predicted by Snell’s law.

Snell’s law is based on the positive refractive indices that are found in common ‘normal’ materials. Veselago considered materials with a ‘negative’ refractive index. Such materials can be made up of tiny structures that are each less than a wavelength of light in size.

This kind of special construction gives ‘metamaterials’ such unusual optical properties. In particular, refraction takes place at the interface between normal and metamaterials such that the light rays bend in the opposite direction with respect to the interface than between two normal materials.

The strange refractive indices of these materials can be engineered to bend light rays in all directions, allowing the incoming rays which would normally be scattered by the object to instead be guided around it. Indeed, British physicist Sir

John Pendry showed that it is possible to build an invisibility cloak using these designer materials.

Another of the unusual properties metamaterials possess is the ability to make perfect images of objects that are very close to a slab of the metamaterial. Even a flat surface is suffcient to make a lens, which makes them suited to viewing very tiny objects—so-called nanostructures, because they have sizes of the order of tens of nanometres. This is the 21st-century version of Hooke’s technology, and will perhaps unleash a similarly fruitful era of discovery.

All of the imaging systems described in this chapter make two-dimensional renderings of objects. That’s normally the way we experience and think about images—as flat pictures. But what if it were possible to conceive of a system that could make three-dimensional images? Remarkably it is, but that requires a deeper view of light itself, which we will consider in Chapter 3.

Chapter 3 Waves

Chapter 2 explored the idea of light as consisting of fundamental particles, moving along well-defined trajectories. This ‘billiard ball’ model, in which a light beam is a collection of individual, compact, and well-localized bundles of energy, stands in contrast to the alternative view, which is that light is a wave. This conception of light has been pursued in parallel with the particle view of light, although it took many years of discussion and experimentation for the wave picture to become fully accepted.

Unexplained phenomena?

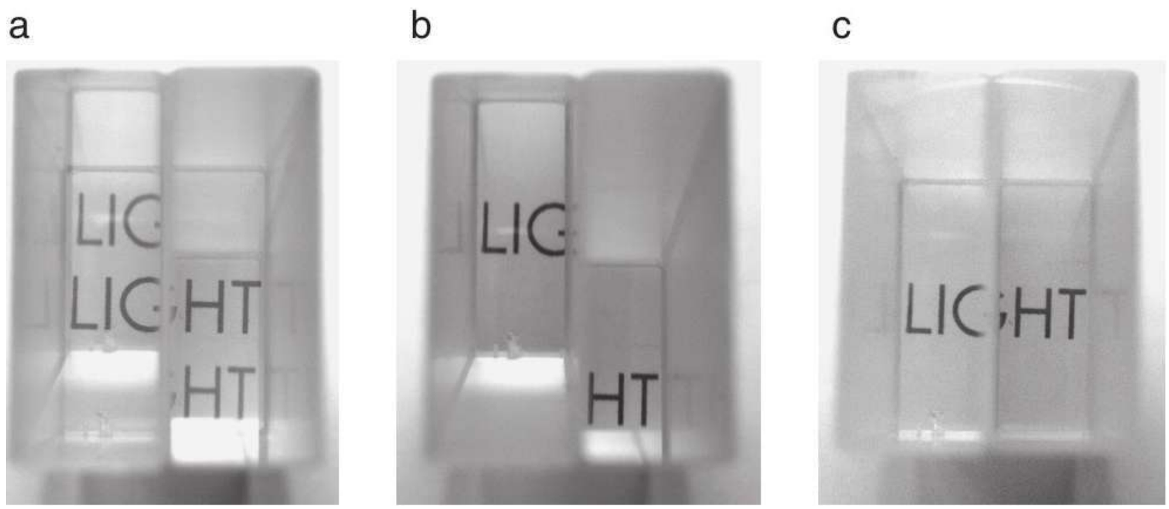
If you look, when the Sun is shining, at a pool of water on which is fioating a thin layer of oil, you will see coloured margins that map out the contours of the oil layer. It was this observation that stimulated historical thinking about light as a wave motion. Newton was one of the first to describe this effect, which provided a challenge to his particle model of light. Newton needed to make severe contortions to his model to account for this observation. Across the Channel, Newton’ s rival in optics and President of the Académie Française, Dutchman Christiaan Huygens, was able to explain this effect using a wave model of light. This turned out to be an altogether more elegant solution. So the

ideas of wave and particle have been around simultaneously since the early days of the scientific study of light.

Other observations, such as those made by Francesco Grimaldi in the mid-17th century, gradually accrued evidence that did not fit with the particle model. Grimaldi saw that light rays deviated from straight lines when they passed through small apertures, such as a tiny hole in a screen. He noted that the light was diffused and that the edges of the beam were fringed with colour, especially pronounced for small objects such as a hair or a piece of gauze. He concluded that the striations seen when light was incident on a small or narrow object were evidence that the light had been bent from its original path as it passed the edges of these objects. If light really consisted of particles moving along straight rays, such solid objects would surely just cast a shadow, and not cause light particles to deviate into strange patterns.

Further, the problem well known to Newton and his contemporaries of the bizarre way in which light was refracted through certain materials, notably crystals such as calcite—a naturally occurring mineral—confounded explanation in terms of simple particles. An example of this behaviour is shown in Figure 15. The word LIGHT, written on a sheet of paper, is illuminated by a light bulb. Two pieces of calcite have been placed over each half of the writing. In the left half of Figure 15(a), two images of the word appear, displaced with respect to each other. In the right half, the

two images are displaced in the opposite direction. The lower image in the left half and the upper in the right are just as one would expect from ordinary refraction of light reflected from the paper, seen through the crystal. But the second set of images appears to arise as if from a different refractive index. By placing a polarizer over the crystal, as shown in Figure 15(b) and 15(c), it is possible to isolate images formed by two different orientations of polarized light. Each polarization experiences a different refractive index. This is the phenomenon of birefringence.



15. An illustration of birefringence. The image of the word LIGHT viewed through a pair of calcite crystals, using a. unpolarized light, b. vertically polarized light, and c. horizontally polarized light.

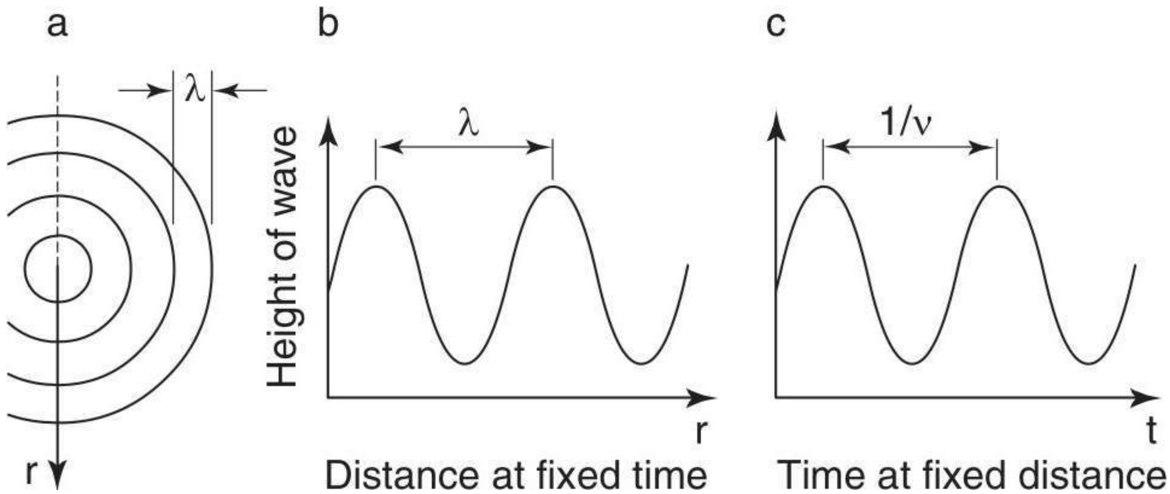
These observations all pointed to features of light that demanded explanation. They are known respectively as interference, diffraction, and polarization. We will explore these

phenomena in this chapter, developing a picture of light as wave motion.

Wavelength and wave frequency

What are the characteristics of wave motion? Waves are a form of undulation associated with a medium, for example water waves on the surface of a pond. These waves consist of the up-and-down motion of water molecules at the interface between the liquid and the air. The highest and lowest points of this motion become the peaks and troughs of the water wave, while the wave itself moves across the surface—that is, at right angles to the motion of the water molecules. For this reason it is called a transverse wave. Its speed depends on the depth of the water, among other things.

The circular surface waves that radiate away from the point at which a stone is cast into the water, as shown in Figure 16a, are a familiar effect. The distance between successive peaks is called the wavelength (Figure 16b), and the rate at which peaks hit the shore is called the frequency (Figure 16c). The product of these two quantities is the speed of the wave.



16. A circular wave on a water surface: a. contours of equal height, called wavefronts, b. wave height vs distance from centre at a particular instant, c. wave height vs time at a particular location.

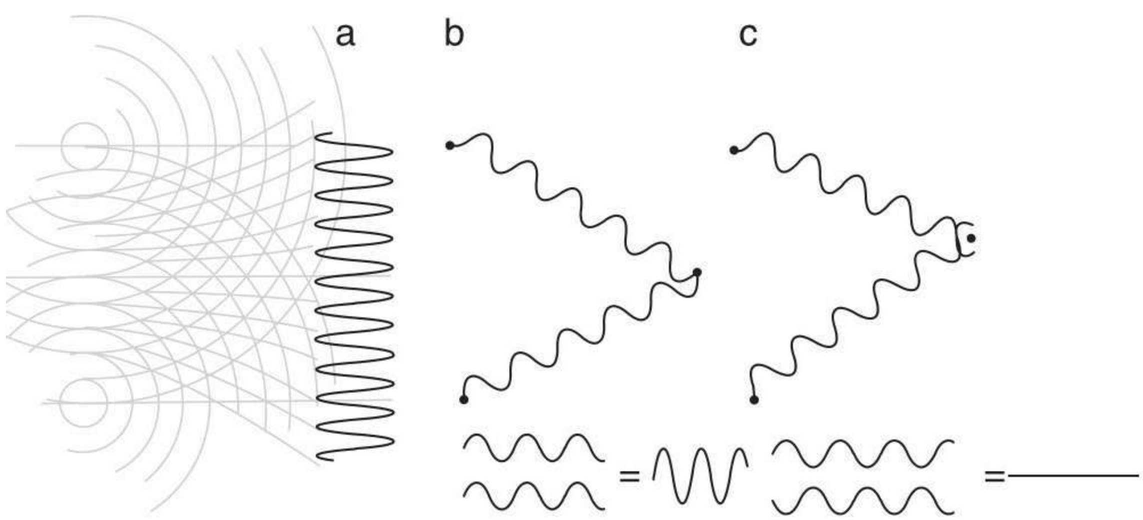
The puzzle for many centuries was what kind of undulations constituted light. It was assumed that some medium would need to exist in order to support a wave. And since the speed of light is so large, this would need to be a very stiff medium. But a stiff medium would surely make it diffcult for other bodies to move through it. For instance, because we can see distant stars, there should surely be some medium to support the propagation of light between the Earth and the star. Since the Earth moves around the Sun, the planet would be continually swept by a ‘wind’ as it travelled through such a medium. This enigmatic medium was called the ‘aether’, and it was not until the end of the 19th century that it was finally discarded as a useful concept.

So, what sort of wave is light? This was also finally answered in the 19th century by James Clerk Maxwell, who showed that it is an oscillation of a new kind of entity: the electromagnetic field. This field is effectively a force that acts on electric charges and magnetic materials. For example, a cloth charged with static electricity will attract dust particles to it. A magnet will be attracted to the door of a refrigerator. In the latter case you can feel this force as you place the magnet close to the door: the magnet accelerates towards the door unless an opposing force is supplied.

In both cases, a force pulls one object to the other, so that at some distance away from the cloth the dust particle feels this force due to the electric field generated by the charges on the cloth. And similarly for the refrigerator door, due to the magnetic field generated by the magnet. In the early 19th century, Michael Faraday had shown the close connections between electric and magnetic fields. Maxwell brought them together, as the electromagnetic force field. It turns out that in the wave model, light can be considered as very high frequency oscillations of the electromagnetic field. One consequence of this idea is that moving electric charges can generate light waves. I’ll discuss this, and other methods for making light, in Chapter 5.

Interference

If two stones are dropped in the water next to one another, the resulting two sets of expanding circular waves collide. Some of the peaks are therefore higher, where the waves reinforce each other. But there are also lines of flat water, where there is no up-and-down movement of the water surface, even though the waves generated by both stones' disturbance of the water are passing through all points along these lines. The location of such lines is shown in Figure 17, along with the wavefronts—the locus of points on the peaks of the waves emanating from each source.



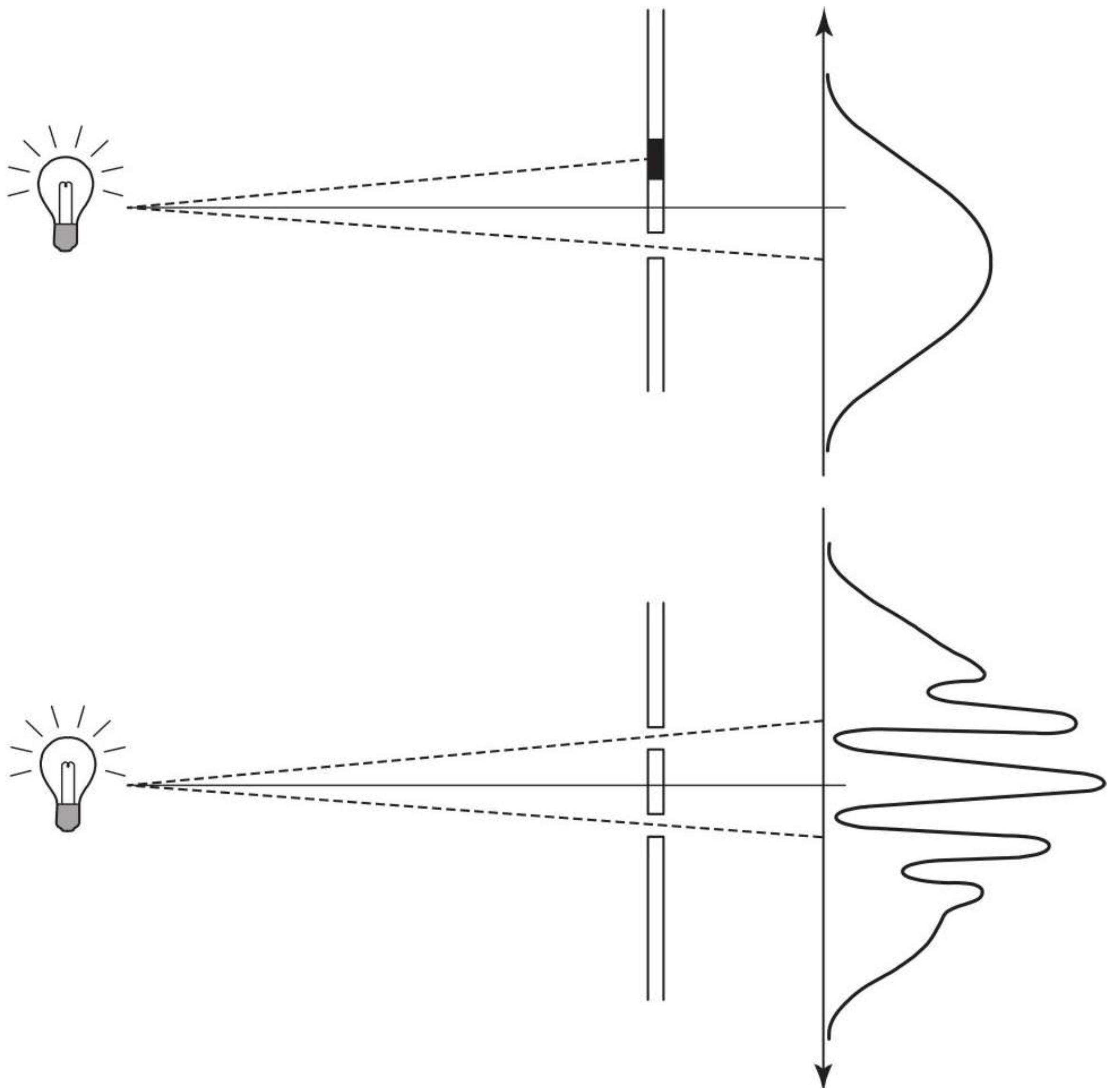
17. a. Two interfering circular waves on a water surface. The dashed lines are contours of equal phase. b. Constructive interference between paths of equal length. c. Destructive interference between paths differing by half a wavelength.

This phenomenon is known as interference, and it arises from the addition of the amplitude of two waves when they meet. If the peaks of the two waves coincide, the resulting

peak is twice the size, or amplitude, of either wave. The waves are said to be in phase, and the interference is constructive. This is shown in the middle diagram of Figure 17. If, however, the waves are exactly out of phase, so that the peaks of one coincide with the troughs of the other, then the resultant wave has an amplitude of zero: the waves have 'cancelled' one another or interfered destructively. This is shown in the bottom diagram of Figure 17. It is immediately clear that such a phenomenon could not happen with particles, for how could two particles cancel one another?

It was the observation of interference effects in a famous experiment by Thomas Young in 1803 that really put the wave picture of light as the leading candidate as an explanation of the nature of light. Young's experiment was very simple, and very elegant. He took a candle as a light source, and placed it behind a screen in which there were two holes a very small distance apart. The light shining through these holes could be seen on a second screen some distance away. Now, if only a single hole was used—say by covering up the second—then the light made a single spot on the screen. When both holes were open, however, something marvellous happened—instead of just a spot of twice the brightness, the spot now had striations. These are generally nearly straight lines of zero intensity perpendicular to the line joining the centre of the two spots, as shown in cross-section in Figure 18 arising from the interference of the waves passing through

the two holes. They are called ‘Young’s fringes’, and are one of the key pieces of evidence for light as a wave motion.



18. Thomas Young’s experiment. The light from one slit produces a smooth intensity pattern. When both slits are open ‘fringes’ appear, characteristic of wave-like behaviour.

How does interference explain the coloured ‘fringes’ observed by Newton in reflections from two surfaces very

close together? What is needed for interference to occur is two waves, the relative phase of which (that is, the relative positions of the peaks of the two waves) can be adjusted. In Newton’s experiment, interference occurs because a single incident beam of light is partially reflected at the two surfaces and thus split into two waves. If the distance between the two surfaces is equal to a single wavelength of the light, then the peaks of the two reflected waves coincide, and there is a bright fringe. If, however, the distance is just half of the wavelength, then the peak of one lines up with the trough of the second and there is a ‘dark fringe’ owing to destructive interference. Thus, when you look at the surface the bright and dark fringes map out the contours of the separation of the surface with a resolution of less than one wavelength. For green light, with a wavelength of approximately one-half of a millionth of a metre, or 500 nm, the separation can be determined to better than 250 nm—a precision of about 1/40th of the diameter of a human hair.

Of course, for a different wavelength the bright and dark fringes will occur in a different place, so that the surface shows fringes of colour when illuminated with white light. It is interference of light waves that causes the colours in a thin film of oil floating on water.

Interference transforms very small distances, on the order of the wavelength of light, into very big changes in light intensity—from no light to four times as bright as the

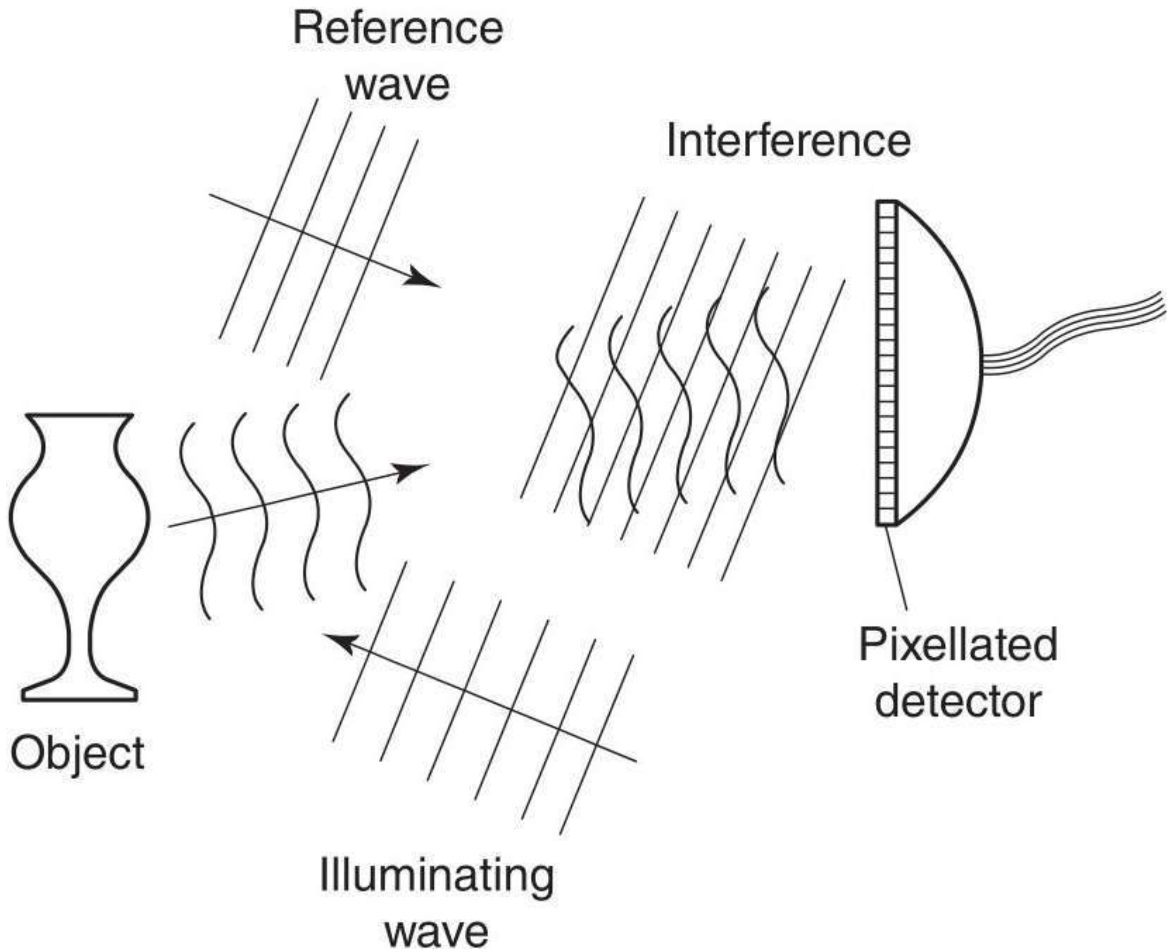
individual constituent waves. Such changes in intensity are easy to detect or see, and thus interference is a very good way to measure small changes in displacement on the scale of the wavelength of light. Many optical sensors are based on interference effects.

Holography

Interference is also the means by which one can make true 3D images, that is, images that can be viewed from different angles and reveal different aspects of the object. These are different to the synthetic images in so-called 3D films, and are called holograms. Holograms are made by recording the full waveform of light scattered from an object. The sort of 2D images we're used to from photography encode only the amplitude of the waves. The phase information is lost. This is because detectors only respond to amplitude, so there is no way in the sorts of images we have looked at previously to extract phase. Nonetheless, it is the phase of the wavefronts scattered from an object that encodes its shape.

What interference enables is the encoding of phase into intensity, so that photodetectors can register patterns in which the full amplitude and phase information of the object wave is recorded. The principle is shown in Figure 19. The wave scattered from the object interferes with a reference wave that has a well-known shape, often generated by a laser. The interference pattern is recorded on a detector or in light-sensitive materials: this is the hologram itself,

invented by Denis Gabor in the middle part of the 20th century.



19. A hologram is constructed by recording the interference fringes between a reference beam of light and one scattered from the object.

Viewing the image is a little more complicated than with an ordinary photograph. First, a reference beam illuminates the hologram and some of its light is scattered from the patterns encoded in the material. These scattered beams have the remarkable property that they reproduce the beams scattered from the original object, so that as your eyes

receive them, it appears that the object is reconstructed in front of you. Moving around the image reveals different aspects of the object, because the beams scattered from those different parts encode different information.

Holograms can also be computer generated and embossed in metals or other materials. The idea is that the surface shape mimics the interference pattern of the reference and object waves, so that raised parts correspond to bright fringes and low parts to dark fringes. Again, illuminating this with a reference wave causes scattered light to imitate the wavefronts of the chosen object. This kind of hologram is used as a security device, including on some banknotes (such as the UK £20 note, which has a strip of holographic images of the 18th century Scottish economist Adam Smith on it), because they are hard to make and require advanced engineering to copy.

Limits to imaging, redux

The wave picture of light also explains why we can't see infinitesimally small objects using a microscope, just as Abbe realized. Very tiny things—down to about half of one micron (one millionth of a metre, or about half a wavelength of visible light) can be seen using an ordinary light microscope. Much more sophisticated methods are required to see even smaller things. The reason is that the wave nature of light puts a lower limit on the size of a spot of light.

I noted previously that when two light beams meet, they interfere to give regions of zero amplitude—dark fringes. The distance between these fringes is actually determined by the angle at which the two beams intersect. If the angle is very large, the fringe spacing is small; if the angle is small, the fringe spacing is larger. The smallest possible separation of the fringes is one wavelength, or about one micron for visible light.

Now, if this fringe pattern is recorded as a hologram, then when it is illuminated again with the reference beam two beams will emerge in the directions of the light beams that were used to record the interference pattern. In order to see such a fringe pattern with a microscope the lens must capture both of these beams to form an image of the fringes. If this didn't happen, say because the lens could only capture one beam, then there would be no interference fringes in the image.

This is the physical basis of Abbe's criterion that I introduced in Chapter 2: the maximum angle between two beams captured by the lens of an imaging system sets the minimum object size that can be resolved. It's straightforward to see then that the best any lens system can do is to see object sizes of about the wavelength of the illumination. Thus conventional light microscopes can see very tiny objects, about fifty times smaller than a human hair, but not much smaller than this. They can be used to look at biological cells, but not the cell nuclei, for instance.

Super-resolution imaging

Optical scientists and engineers have found a number of very clever ways to get around the object size limits posed by conventional microscopes, so that they can see inside cells, or view objects that are more than one hundred times smaller than the wavelength of light. These instruments make use of new materials and processes, such as the ability to attach nanometre-scale particles to the objects you wish to see, or to insert molecules that emit light into cells. These emit light (they fluoresce) at a long wavelength when they are illuminated by a beam of light at shorter wavelengths. Since they are much smaller than the resolution of the microscope lens, the resulting image is a spot the size of which is entirely limited by the microscope optics, according to Abbe’s formula. But the exact centre of this image can be located very precisely by using a camera to take a long look at the fluorescence of the attached nanoparticles and determine the position of the maximum intensity of the spot. This approach is called photo-activated localization microscopy, or PALM, and was invented by Eric Betzig in the US. It has revolutionized live cell imaging, allowing faster acquisition and precise depth resolution over a wide field of view.

Another method for measuring small structures in larger fluorescent objects is to shine a second annular beam on them that causes the objects illuminated by this beam to have their fluorescence extinguished, so that the remaining

fluorescence can be located precisely by the same approach as described previously. This method is called stimulated emission depletion microscopy, or STED, and was invented by Stefan Hell in Germany. I will describe the process of stimulated emission in more detail in Chapter 5. These novel ways of making high-resolution images that enable intracellular structures to be imaged have had enormous impact in biology and medicine. Indeed the significance of this impact was recognized when the 2014 Nobel Prize in Chemistry was awarded to Betzig and Hell.

The process identified by Abbe works in reverse, too. When applied to the illumination of a sample by means of a microscope lens, it says that a light beam cannot be focused to a spot much smaller in diameter than a wavelength. Again, the tightness of the focus depends on the range of angles that the lens can produce on the side that faces towards the object: the broader the range of ray directions, the tighter the focus of the light.

The relationship between the range of angles between interfering beams and the size of the fringe structures turns out to be a very fundamental property of waves. This idea was quantified by Joseph Fourier, a French scientist of the early 19th century, who provided a detailed mathematical analysis of light wave propagation. Fourier's theorem says in simple terms that the smaller you focus light, the broader the range of wave directions you need to achieve this spot.

Diffraction

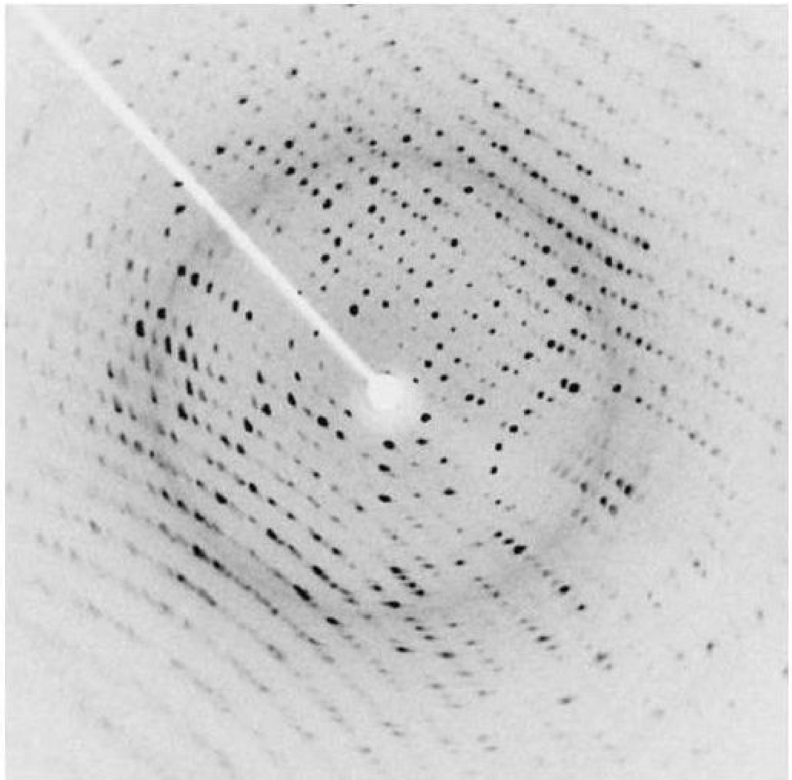
This explains another feature of light beams—they gradually diverge as they propagate. This is because a beam of light, which by definition has a limited spatial extent, must be made up of waves that propagate in more than one direction. The idea can be tested using a laser pointer. The beam emitted by the laser itself is about 10 μm (ten millionths of a metre) or so in diameter. When it reaches the screen, it is about 1 mm (a thousandth of a metre) in diameter. And if it were sent further away, say to the Moon (approximately 400,000 km), then it would be about 24 km in diameter! This phenomenon is called diffraction.

Diffraction has some interesting applications in determining the shape and symmetry of structures. For instance, when you shine a beam of light on a screen with small holes in it, of diameter comparable to the wavelength of the light, the light diffracts through the apertures, spreading in inverse proportion to the size of the aperture. These diffracted beams interfere some distance away from the screen, and the resulting interference fringes—the so-called diffraction pattern—tell something about the size and relative location of the apertures. For instance, if they are in a regular array, then the diffraction pattern will also show regularities. The advantage of using such patterns to measure this type of object is that you don't need to have very expensive or complicated lens systems or detectors close

to the object—you simply look at the pattern when it has naturally expanded due to diffraction.

Now, imagine that the screen is replaced by a transparent solid material, say a crystallized protein structure. The ‘holes’ are replaced by the atoms in the protein molecule, which are very small indeed and are connected to one another by bonds in the molecule that are about one tenth of a billionth of a metre (0.1 nm) in length.

If light with a wavelength of about this size illuminates such a structure, then the light will be diffracted. The actual structure of the molecule itself can be determined from the diffraction pattern. This is the basis of X-ray diffraction. As noted in Chapter 1, it was famously used in the effort to find the structure of DNA and is now a very common tool in biochemistry, used regularly for finding out the structure of new molecules that might be useful in developing drugs for example. It requires a bright X-ray light source, as well as a means to make crystals out of the molecules. Figure 20 shows a diffraction pattern from a crystal of bovine enterovirus.



20. An X-ray diffraction pattern of a protein crystal taken using a modern synchrotron X-ray source.

It’s clear that if you want to transmit light over long distances, then diffraction could be a problem. It will cause the energy in the light beam to spread out, so that you would need a bigger and bigger optical system and detector to capture all of it. This is important for telecommunications, since nearly all of the information transmitted over long-distance communications links is encoded on to light beams.

Guided waves

The means to manage diffraction so that long-distance communication is possible is to use wave guides, such as

optical fibres. A wave guide is a structure that has a carefully designed profile for the refractive index. For example, the index variation across an optical fibre is arranged so that the ‘core’, with a diameter of a few millionths of a metre, has a higher index than the surrounding ‘cladding’. The light is confined to this higher index core and moves along the fibre without diffracting, remaining the same size over very long distances—across the Atlantic Ocean via undersea cables, for example. This means of controlling light is the basis of a wide-ranging optical information infrastructure, from communications to sensors.

Polarization

The final important feature of the wave model is the characteristic of polarization. Recall that in transverse waves the undulations occur in a direction at right angles to the direction of propagation of the wave. Importantly there are two such directions for these undulations.

Consider waves on a string, for instance. If you move one end of a rope up and down rapidly, you can see the undulations move along the rope. A similar thing happens if you move the end left and right equally rapidly. Both vertical and horizontal oscillations are at right angles to the motion of the wave down the rope. The waves are said to be ‘transverse’.

Light waves behave similarly. For example, horizontally polarized light has an electric field that oscillates in a horizontal plane (relative to, say, an optical bench). Similarly, the field of a vertically polarized light beam oscillates in a vertical plane. (There are more complicated forms of polarization too, but these will sufice for now.) The phenomenon of birefringence can be explained by noting that a crystal is a structure of atoms in very specific and highly repeatable units. These 'unit cells', consisting of a few atoms, may themselves be asymmetric. Depending on whether the light is polarized along the long axis or the short axis of the unit cell, it will experience a different refractive index, and thus will be deviated by different amounts as it propagates through a block of such material.

One of the well-known ways in which polarization of light is exploited is in sunglasses. Some of these use lenses made of plastic material (for example, a sheet of Polaroid plastic) that acts as a polarizer—that is, an element which transmits light of a particular polarization (say vertical) and absorbs that of the orthogonal polarization (horizontal). Polaroid is made from rugby-ball shaped molecules that are aligned and 'frozen' in the plastic polymer. These molecules preferentially absorb light that is polarized along the axis of alignment of the molecules. Since generally sunlight scattered from objects has a random polarization (roughly 50 per cent of each polarization direction), then filtering out light of one polarization effectively reduces the brightness of the scene by half. Further, polarized

sunglasses reduce glare: that is, light reflected from shiny flat surfaces, such as the hood and windscreen of cars. These surfaces tend to reffect a higher proportion of light polarized in a direction parallel to the surface (an effect discovered in the 19th century by Sir David Brewster and now named for him). Such reflections are blocked by spectacles that are constructed in the manner noted here, making for clearer vision of the road.

Transparent birefringent materials can also change the polarization of light without absorbing the light. This is because the speed of light depends on the direction of polarization with respect to the 'orientation' of the material. Some materials, such as ordinary glass, have no special orientation: you can rotate the material with no change in the effect on a light beam. As noted above, atoms in birefringent materials are arranged in such a way that there is a preferential direction—the symmetry axis—along which atoms respond differently to light. That is, light polarized along the symmetry axis will go slower (say) than light polarized perpendicular to the axis. Now, imagine that the light is polarized at 45 degrees to the symmetry axis. We can consider that half of this light is then polarized along the symmetry direction and half polarized at right angles to this direction. If the latter is slowed down by a suffcient amount, then the light emerging at the other face of the material can be polarized at -45 degrees. Thus the direction of polarization has been 'rotated' by 90 degrees.

Some birefringent materials can be used to control the state of polarization, by actively adjusting the orientation of the alignment axis of the molecules, say, using an applied voltage across the material itself. An example of this is a class of materials called liquid crystals (LCs), which consist of elongated molecules. The orientation of the molecules in an LC can be controlled by applying a voltage across them. Other materials become birefringent when forces or stresses are applied, because the force causes the molecules to rotate or the atoms to change alignment. This phenomenon enables the construction of force sensors by monitoring the state of polarization of light at the output of the optical sensor.

Placing a piece of birefringent LC between two polarizers also allows control of the light intensity by means of an electrical connection. Applying a voltage reorients the molecules, thus changing the refractive index seen by a polarized light beam. If a polarizer is placed after the LC then depending on the applied voltage a greater or lesser amount of light will be transmitted through the final polarizer. An array of such 'cells', each driven by a separate electrical signal, can be used to form a display where each cell is a single pixel. This is the basis of a liquid crystal display (LCD), and is often used for computer display screens and television sets.

In fact, such displays can be used to show 3D movies. The illusion of depth in these movies derives from the

stereography of human vision. Each of our eyes sees a scene from a slightly different location, since they are a few centimetres apart in our skulls. The two images are combined in our brains to give us a perception of depth.

This illusion is reproduced using 3D glasses by means of polarization. Two images are projected on to the display or screen. Each of these is generated using light of a specific polarization, and each is shot from a slightly different vantage point. The 3D glasses are polarizers set at different orientations that allow transmission of one scene to the left eye and one to the right, each completely blocking the alternate image that has the 'wrong' polarization. Thus we have a sense that the scene is as we would perceive it in the natural world—that is, the illusion of a three-dimensional arrangement of objects and people.

The successes of the wave model of light have been stunning, allowing us to understand some important characteristics of light and to use this understanding to build new technologies. And the successes of the ray picture of light have been equally amazing. Yet it is certainly puzzling that two apparently very different views of what light is should be necessary. It is to this conundrum that I now turn.

Chapter 4 Duality

The two different views of light, as a particle and as a wave, both contain insight and value. They have each enabled both new understanding of the natural world and the development and design of new technologies. Yet they appear to be vastly different in their conception of what light actually is. On the one hand, the particle model views light as a localized entity, a bundle of energy, that moves along a well-defined trajectory. On the other hand, the wave model describes light as a diffuse entity, permeating through space with no connection to the motion of solid things. How can these two pictures possibly refer to the same thing? This dilemma was recognized early on by Huygens and his contemporaries, but the two views remained in tension, as alternative descriptions of light, until the 19th century.

When Maxwell developed his theory of electromagnetic fields, he was able to use this to explain the properties of light as wave motion of those fields, as we saw in Chapter 3. This triumph of reasoning appeared to confirm the experiments of Thomas Young and Auguste Fresnel (described in Chapter 3) by providing an explanation of two fundamental phenomena, interference and diffraction, that did not easily fit within the particle model. Yet the concept of trajectories remained, and still remains, an extraordinarily powerful one for the analysis and design of optical systems. So there’s an uneasy truce of these two pictures—a dualism within classical

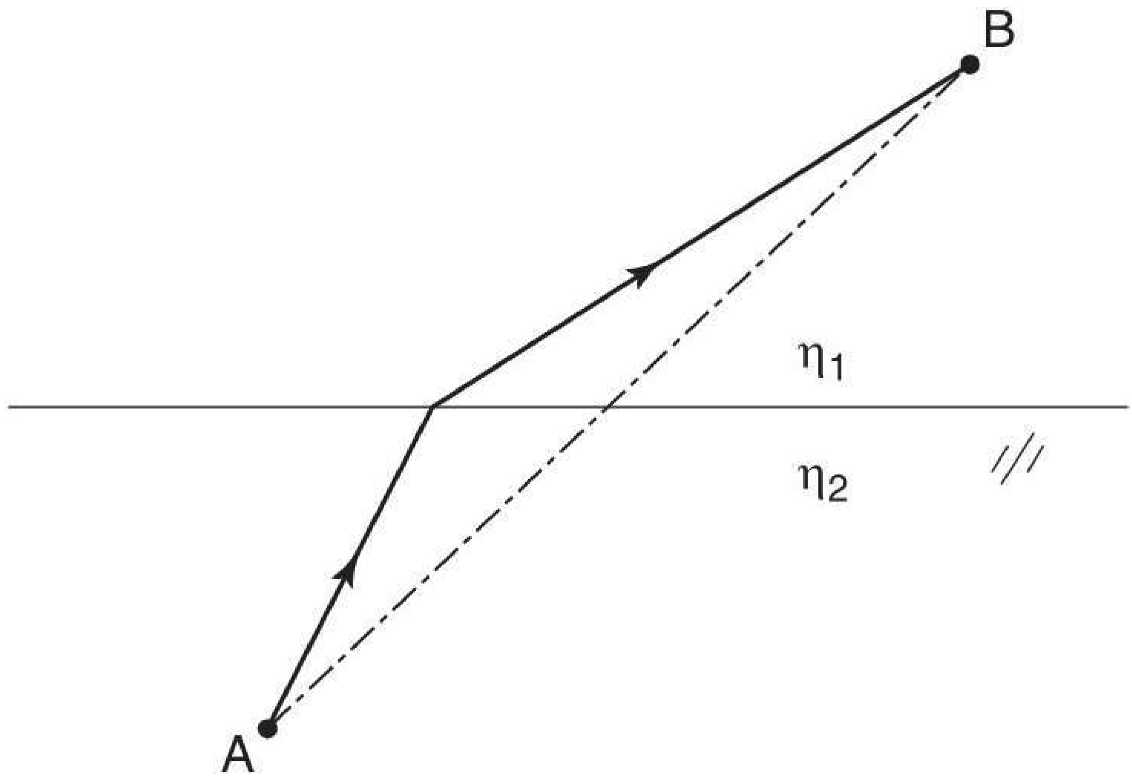
physics—that requires some consideration. How can they be reconciled?

Looking at trajectories again

In the 17th century the Frenchman Pierre de Fermat proposed an ingenious formulation of refraction that was very different from that of Snell. Recall that Snell's law deals with the change of direction of a ray of light at an interface between two transparent media. The ray, defined by the direction in which it is travelling towards the interface and the point at which it hits the interface, has its direction altered by an amount proportional to the ratio of the refractive indices of the two materials. It is only the local properties of the ray and interface that are important. Snell's law applies at each point along the trajectory, as if the ray is 'feeling' its way along, adjusting direction when it encounters a new interface.

Fermat's conception was radically different. He argued that one should define the trajectory in terms of starting and ending points, as shown in Figure 21. He suggested that the question to ask is: what is the path that the light takes to traverse the space between the two points? He proposed that it should take the path that minimizes the time of flight between the two points. That this gives the same answer as Snell is remarkable and profound. Fermat's 'principle of least time' suggests that the light considers the overall picture of the situation, and that the notion of

a ray is one that takes into account both the initial and final positions and directions as well as everything in between. The contrast with the local model of a particle reacting to its immediate environment is telling.



21. Fermat’s conception of a light ray as a path of least time connecting the start and end of the trajectory. The ray crosses an interface between two optical media in which light moves at different speeds.

This idea was taken up by the German natural philosopher Gottfried Wilhelm von Leibniz, who was Newton’s contemporary and antagonist. Leibniz was impressed by the holistic picture of the process described by Fermat, and the concept of ‘optimization’ that it implied: a ray explores the whole of space and picks just that path that will minimize its transit

time between the specified beginning and end points. He developed the mathematical tools for analysing this idea—the calculus of variations—by which the effects small changes on a trajectory would have on the time taken to traverse the modified trajectory could be calculated. Leibniz recognized the importance of the notion that Fermat’s principle provided: the movement of light from one point to another defines an ‘optimal’ trajectory.

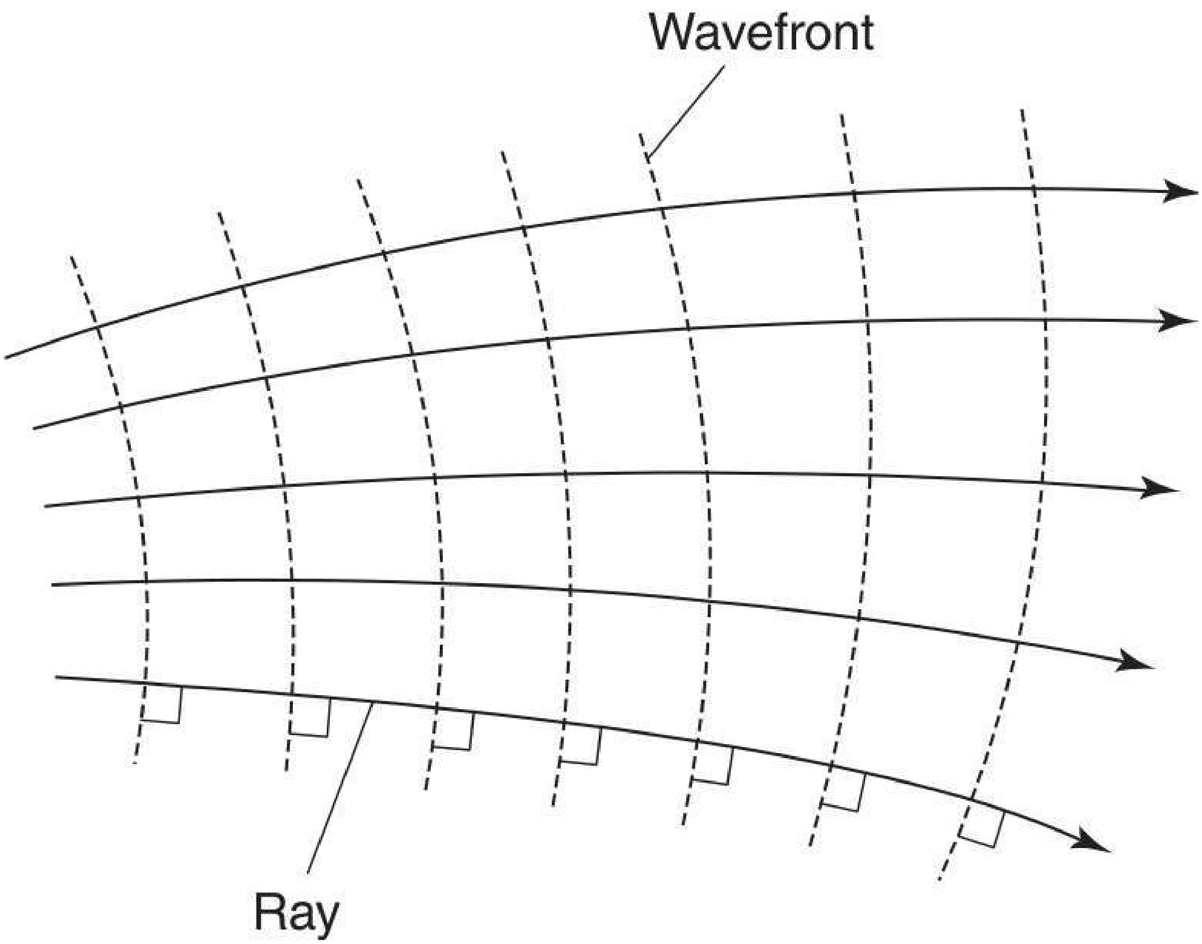
Indeed, so taken was Leibniz by this concept of optimization that he elevated it to a teleological principle: that the world itself, in all its aspects, was on the optimal trajectory between a starting point and a finishing point. The contradictions inherent in such a position, when applied outside of the realm of science, were ably lampooned by Voltaire in his novel Candide, where Leibniz’s ideas are put into the mouth of Dr Pangloss, who insists disasters both natural and man-made were nonetheless evidence that this is the ‘best of all possible worlds’.

Connecting waves and rays

Nonetheless, Leibniz’s mathematical ideas proved to be very fruitful. They were taken up by the renowned Irish mathematician William Rowan Hamilton in the 19th century. He showed formally how the idea of a wave can be allied to that of a collection of particles. Waves can be defined by their wavelength, amplitude,

and phase (see Figure 15). Particles are defined by their position and direction of travel (see Figure 5), and a collection of particles by their density (i.e. the number of them at a given position) and range of directions. The media in which the light moves are characterized by their refractive indices. This can vary across space. For example, at the interface shown in Figure 20 there is a step change in the refractive index across the boundary between the two media.

Hamilton showed that what was important was how rapidly the refractive index changed in space compared with the length of an optical wave. That is, if the changes in index took place on a scale of close to a wavelength, then the wave character of light was evident. If it varied more smoothly and very slowly in space then the particle picture provided an adequate description.He showed how the simpler ray picture emerges from the more complex wave picture in certain commonly encountered situations. The appearance of wave-like phenomena, such as diffraction and interference, occurs when the size scales of the wavelength of light and the structures in which it propagates are similar. Thus you see diffraction patterns arising when the object that the light hits is a few microns in diameter, or has a very sharp edge, such as the delicate structure in a bird’s feather, or a butterfly wing. Otherwise, as in the case of a camera lens, the trajectory provides a suffcient description, since the refractive index is uniform throughout the glass of the lens itself.



22. Hamilton’s idea of rays as connecting wavefronts—thus joining the two primary conceptions of light.

Further, Hamilton showed that Fermat’s trajectories related directly to a property of the wave—the wavefronts. These are the loci of points at which the wave has the same phase at each point in space. For instance, when you see the ripples on the surface of a pond after a stone has been cast into it, the circular patterns are just these wavefronts. They are the places on the surface where the wave ‘peaks’ (or troughs) at a given instant of time. Now, what Hamilton noted was that rays could be considered as lines that

intersected the wavefronts at right angles, as shown in Figure 22, thus connecting adjacent wavefronts by a well-defined trajectory.

Hamilton’s ‘optical analogy’

This remarkable result suggested another profound comparison—Hamilton’s so-called ‘optical analogy’. What he noted was that the well-known formulation of mechanics—the motion and position of solid bodies of matter—was based on the idea of trajectories. The idea that these may also be in some sense ‘optimal’ had been considered by Pierre Louis Maupertuis in the 18th century.

Maupertuis had formulated a way to evaluate the optimal value of a quantity called the ‘action’—essentially the velocity of the body multiplied by the distance it moves (and times its mass)—along the body’s trajectory.

He argued that the action should be minimal for an actual trajectory between two points, just as in Fermat’s argument that the time taken to traverse a light ray should be minimal. Maupertuis’ ‘principle of least action’ is very similar in concept to Fermat’s ‘principle of least time’. Indeed, Leonhard Euler, a Swiss mathematician, showed how to use Leibniz’s calculus to derive Newton’s famous equations of motion from Maupertuis’ principle. Thus Euler connected a description of a trajectory in terms of a particle sensing its way through its environment to one in which the whole of

space between the specified starting and finishing points influences the path.

What Hamilton did was to find equations that encapsulated the variations in action in terms of a simple description of the specific environment in which the body was moving. And this equation turns out to have a very similar form to the one he found for describing the trajectories of light rays (for which the environmental description is just how the refractive index changes with position in the medium). So there is a hint of a latent analogy between the trajectories of solid objects and a fictive wavefront: perhaps all bodies might have both particle-like trajectories and wave-like properties? Indeed, Hamilton's equation, and his eponymous function, turns out to be very important in thinking about the next big step in understanding light—quantum mechanics.

Unsolved puzzles

This was by no means the only hint of a new opportunity for science. About this time, towards the end of the 19th century, light still offered a few puzzles that were unexplainable in terms of the prevailing models of its properties, even with the reconciliation that Hamilton had provided. Two of the most important of these were: the colour of hot objects (including the Sun), and the colour of different atoms in a flame.

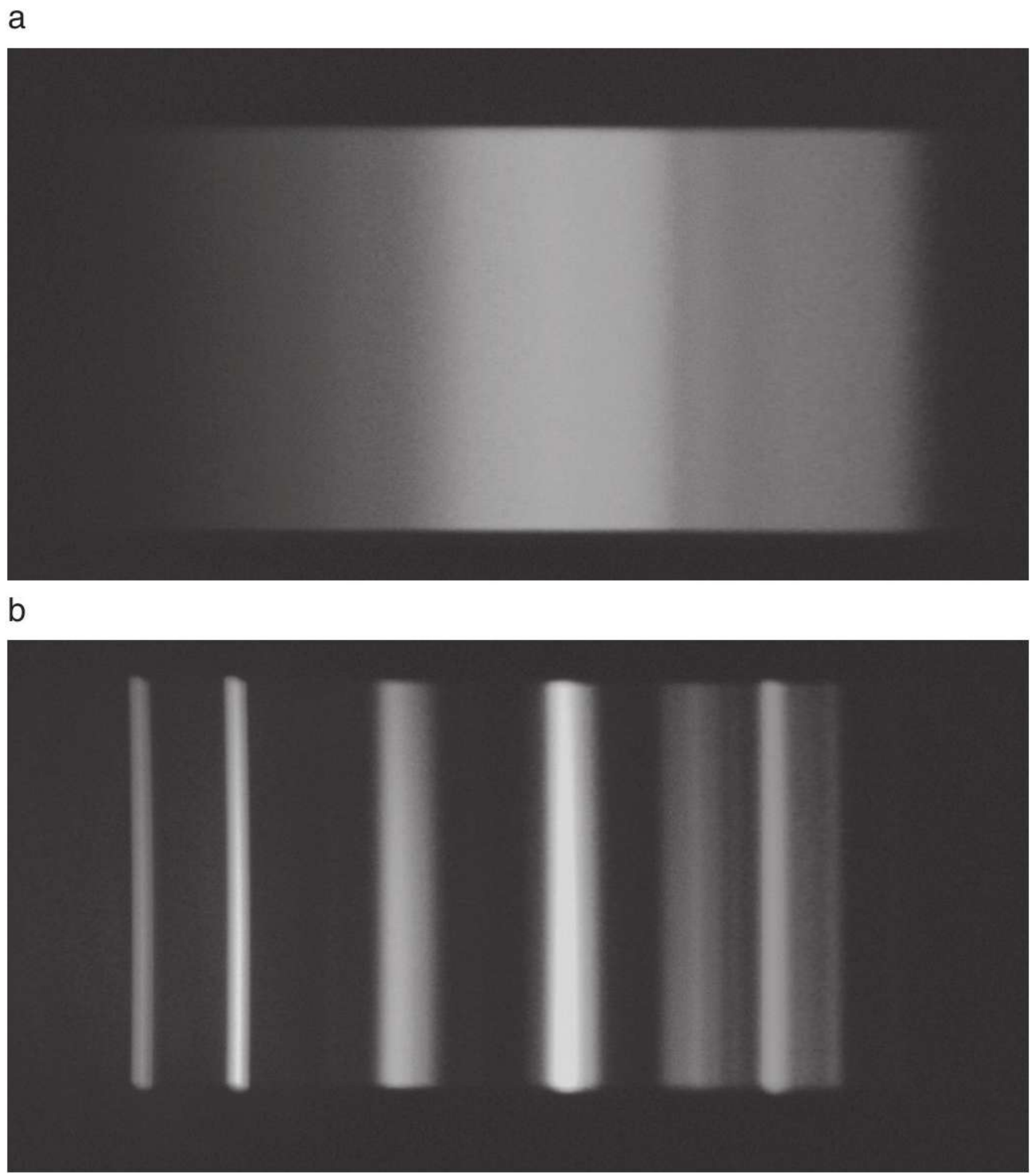
When things are heated up, they change colour. Take a lump of metal. As it gets hotter and hotter it first glows

red, then orange, and then white. Why does this happen? This question stumped many of the great scientists of the time, including Maxwell himself. The problem was that Maxwell's theory of light, when applied to this problem, indicated that the colour should get bluer and bluer as the temperature increased, without a limit, eventually moving out of the range of human vision into the ultraviolet—beyond blue—region of the spectrum. But this does not happen in practice.

The second example arose from the study of light emitted by atoms, pioneered by a Swiss schoolmaster, Johannes Balmer. We'll look at this mechanism in more detail in Chapter 5, but the important feature of the light is in the distribution of colours in its spectrum, shown in Figure 23a. In this respect it is very different from sunlight (the Sun is a good example of a hot body), which has the familiar 'rainbow' spectrum, shown in Figure 23b, consisting of all colours continuously from red to violet (and beyond at each end—just not visible to us). By contrast, a collection of atoms emits a discrete set of colours—a set of spectral 'lines' of particular wavelengths—specifically associated with the internal structure of the particular atom involved.

Both of these phenomena required a radical revision of thinking about light, because they could not be explained within the contemporary models of wave motion and atomic structure.

Max Planck, working at Humboldt University in Berlin in the late 19th century, first came up with an idea to explain the spectrum emitted by hot objects—so-called 'black bodies'. He conjectured that when light and matter interact, they do so only by exchanging discrete 'packets', or quanta, or energy. Planck recognized the revolutionary nature of his idea, and was therefore reluctant to infer too much about light itself from it, though it would eventually dramatically change our view of what light is. His notion revived the idea of light as a stream of particles—discrete objects that carried a fixed amount of energy that could be absorbed by atoms or emitted by them.



23. Spectrum of light emitted from a. the Sun (a ‘black body’), and b. a neon lamp. The former has continuous band of colours, whereas latter shows the discrete lines of particular colours that are a ‘fingerprint’ of the neon atoms.

This seemed like a retrograde step: the wave model of light could explain all of the hitherto observed effects, and it was clear from Hamilton's work that even trajectory-like behaviour, previously the most obvious evidence of a particle-like entity, emerged from a wave model of light in certain common situations. So the idea of a particle of light appeared not to be necessary. Surely it was simply a calculational 'fix', thought up to get out of a tight spot, and would eventually be replaced by a more reasonably consistent picture of light. However, combined with Balmer's observations, this conjecture was set to radically change physics.

In the years immediately after Planck's suggestion, Albert Einstein used the idea of discrete exchanges of energy between light and matter to expose another piece of physics that had eluded explanation—the photoelectric effect. This effect is seen when light shines on a piece of metal. Some electric charges—electrons, in fact—are ejected from the metal. The speed with which the electrons are ejected depends on the wavelength of the light. The light must be suffciently 'blue', that is have a short enough wavelength, in order to see any electrons emitted at all. As it gets bluer and bluer, the electrons are ejected with more and more energy, and thus higher and higher speed.

Einstein explained this by noting that the electrons required a minimum amount of energy to escape the clutches of the metal, and by assigning discrete amounts of energy to a

particle of light—the photon—proportional to the frequency of the light (the constant of proportionality being known as Planck's constant, h), so that when the frequency of the light is high enough (and thus the wavelength short enough) the light, when it is absorbed, can provide enough energy for an electron to escape. His model suggested that the origin of the discrete exchange of energy between light and matter that is central to Planck's notion arises from the actual discrete character of light—a full revival of the particle model.

This idea chimes nicely with Balmer's observations of discrete line spectra of light emission from atoms. But a full explanation of this phenomenon clearly requires a bit more thought about why atoms would deliver light in such packets. The key idea came from Niels Bohr, a Danish physicist working in Manchester. He suggested that the reason light was emitted as discrete packets of energy was that the atoms themselves could only exist in certain configurations. He imagined atoms as analogous to tiny planetary systems: electrons in orbit around a central nucleus. The electron can 'jump' between two stable orbits, emitting or absorbing light as it does so (depending on whether the jump is to an orbit of lower or higher energy). The character of these orbits—or quantum states—depends on the details of the atom: how many electrons and the size of the nucleus. Therefore the energy given or taken up when an electron moves between two quantum states is a signature of the particular atom itself. So when the energy of a light particle—or

photon—matches that of the difference in energy of two quantum states of the electron in an atom, then absorption or emission is possible. Bohr’s ideas explained Balmer’s observations neatly, and added some weight to the idea of light beams as a collection of discrete particles.

All these developments had the potential to undermine the wave model of light that had been so strongly affrmed by Maxwell’s theory. They went beyond even Hamilton’s reconciliation of trajectories and wave motion, since they appeared to be fundamental, not simply the result of an approximation about size and scale. Thus they reopened the question of the nature of light once again.

In 1908, George Taylor, working in Cambridge, performed Young’s double-slit experiment with exceptionally feeble light—so weak that on average there was less than one photon in the apparatus at any time. Yet he still saw interference fringes. That outcome is very strange. If we think of one path from the light source to the detector as being via one slit, and a second path via the other slit, then there are two ways that the photon can get from source to detector. The evidence of interference fringes posed a dilemma that caught the attention of the leading scientists of the day. Bohr captured the diffculty, noting that we would ‘be obliged to say, on the one hand, that the photon always chooses one of the two ways, and on the other that it behaves as if it passed through both’. Even single particles can exhibit wave-like behaviour.

Waving altogether

As you might imagine, it took a truly radical thought to figure a way out of this conundrum. That thought occurred to Paul Dirac, a physicist working at Cambridge in the 1920s. Dirac suggested that the fundamental property of light was that it was both a particle and a wave at the same time. Now, this might appear to you as simply sophistry, a logical trick that resolves nothing. But there is much more behind it than a slogan. What Dirac did was to develop a quantum mechanical version of Maxwell’s theory of electromagnetic fields. Using this he was able to show that, if you measured these ‘quantum fields’ using a set-up like Young’s double-slit interferometer, you would see interference effects characteristic of wave-like behaviour. Whereas, if you simply measured the intensity of the light, you would be eflectively counting the number of photons in the beam.

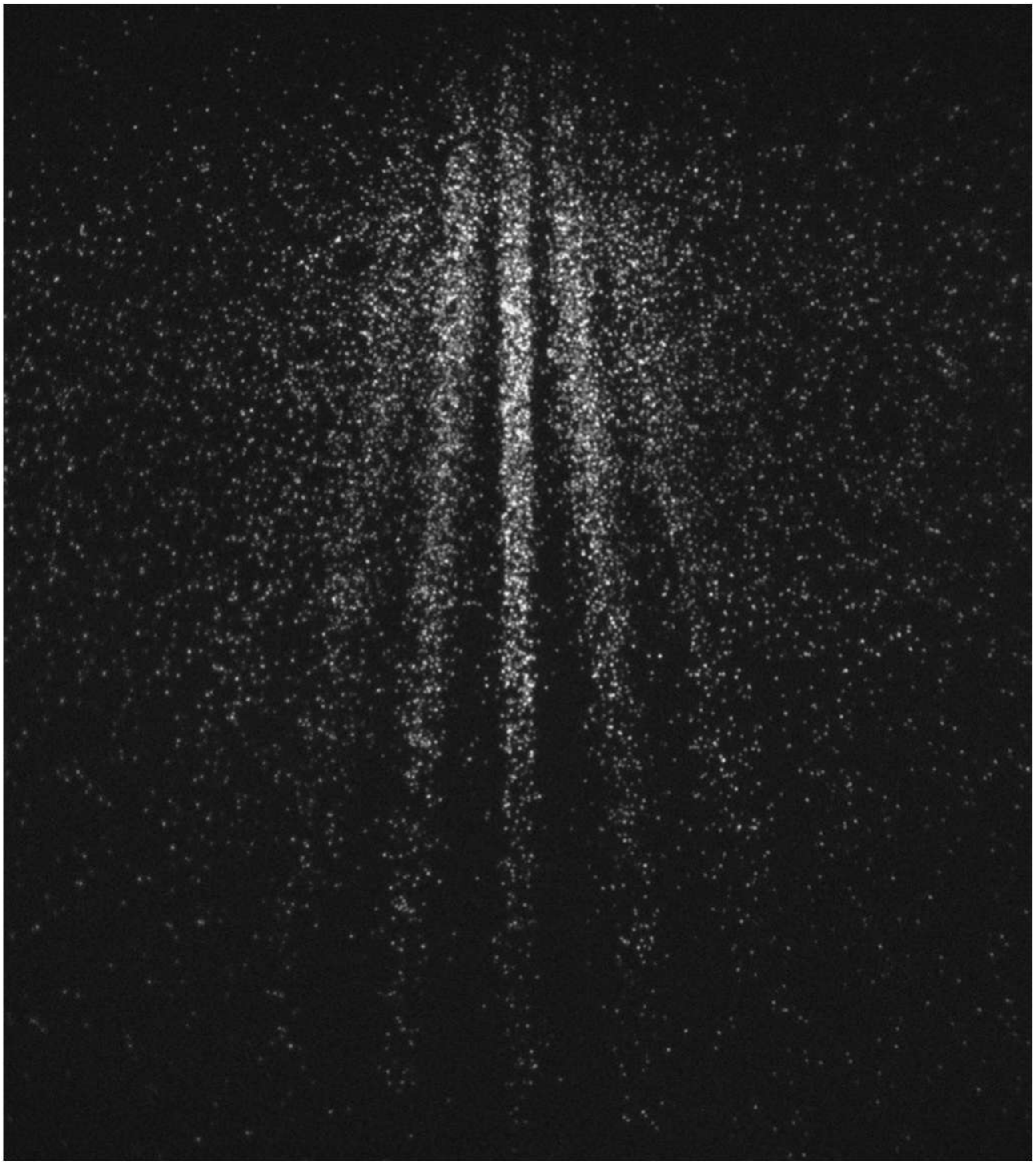
This turned out to be a very profound step. It set the quantum field up as the fundamental entity on which the universe is built—neither particle nor wave, but both at once; complete wave- particle duality. It is a beautiful reconciliation of all the phenomena that light exhibits, and provides a framework in which to understand all optical effects, both those from the classical world of Newton, Maxwell, and Hamilton and those of the quantum world of Planck, Einstein, and Bohr. But the cost is a perplexing and non-intuitive entity at the heart of nature—a quantum field—of which light is but one example.

The radical idea that light was both wave and particle stimulated some major new ideas. For instance, Louis de Broglie suggested that if this kind of dualism existed for light, surely it should for all other things too. So, material bodies that we normally had considered only as collections of particles should also have 'wave-like' properties, taking the next step beyond what Hamilton had considered. He even defined what the wavelength should be (now called the de Broglie wavelength λ_{dB})—proportional to the inverse of the particle's momentum (i.e. its mass and velocity, the constant of proportionality again being Planck's constant):

$$\lambda_{dB}=h/mv$$

This suggests that to look for such effects you should use either very light particles or very cold ones that are moving very, very slowly. Such effects can be observed. Figure 24 shows an interference pattern made using a double-slit-like interferometer, but using molecules instead of light. The implications of this are mind-boggling. If you think of a molecule as just a very light particle, then you cannot explain the pattern, because you consider that it must have passed through one slit or the other. However, the idea that a particle with mass could be so delocalized as to have passed effectively through both slits to interfere with itself is astounding and beggars belief.

If material objects also behave as waves, then, Erwin Schrödinger conjectured, surely there must be a wave equation describing their behaviour. Where to start looking for such a thing? He took the equation that Hamilton had developed, by means of his optical analogy, to describe how the 'action' of a particle evolved. A simple addition to this equation, involving Planck's constant, moulded it into one that described wave motion. This was the origin of Schrödinger's famous 'wave function'. The wave function has a number of properties very analogous to optical waves, including interference and diffraction, but nonetheless refers to things that in the language of pre-quantum physics are palpably particles possessing mass and weight. As a consequence, there are still questions as to what the wave function actually describes. Is it the actual particle itself, or is it a sort of shorthand for what we know about the particle?



24. An interference pattern made using molecules passing one at a time through a tiny version of Young’s apparatus—two very small slits separated by billionths of a metre.

An important quantity for light is its intensity, proportional to the square of the amplitude of the field.

This is related directly to the density of photons in the light beam. Similarly, the square of the wave function is related to the density of particles at a particular point in space at a particular time. But it is impossible to say for certain that a particle occupies a specific position at a specific instant. This indeterminacy seems to be a very fundamental property of the world, and it is related deeply to the fact that quantum fields lie at the heart of things.

Nothing is something

Another consequence of this fact is that ‘nothing’ is actually something. That is, the complete absence of matter (e.g. electrons or atoms) or light (i.e. photons) still has measurable properties. This void is called the ‘electromagnetic quantum vacuum’, and is the state of the universe from which all extractable energy has been removed. Yet it is a seething mass of activity, consisting of fluctuating fields but containing no photons whatsoever. What is even more surprising is that the quantum vacuum has observable consequences. How can this ‘nothing’ give rise to an effect we can see?

We’ve seen that light can be thought of as a wave motion of the electromagnetic field. Imagine this field as ripples on the surface of the sea. These would buffet about any boat, but would not move the boat up and down or push it along as a well-defined wave might. On average, the boat does not move, but it nonetheless rocks back and forth. Now imagine the same

thing for a charged particle like an electron. It ‘feels’ the random changes in the electromagnetic vacuum, being buffeted by these. If the electron is bound in an atom, this buffeting is revealed as a shift of the energy of the quantum states that the electron may occupy. Since the frequency of the photon that the atom may absorb depends on the difference in energy of two such states, these changes can be seen in a change of the colour of light that the atom may absorb. The changes are tiny—a shift of less than one billionth of the wavelength of the light—but measurement techniques for frequencies are so precise that such changes can indeed be determined. The first person to do this, Willis Lamb, working in New York in the 1950s, won a Nobel Prize for showing this frequency shift, which is now named for him.

The dual identity of light has numerous facets. Even in the pre-quantum world, the dichotomy of ray and wave demanded a resolution. That came about by understanding the nature of the wave motion that light embodied, and the scale and nature of objects with which it interacted. Particle-like behaviour—motion along a well-defined trajectory—is suffcient to describe the situation when all objects are much bigger than the wavelength of light, and have no sharp edges. Quantum mechanics puts a new twist on this duality. Light acts as a particle of more or less well-defined energy when it interacts with matter. Yet it retains its ability to exhibit wave-like phenomena at the same time. The resolution is a new concept: the quantum field. Light particles—photons—are

excitations of this field, which propagates according to quantum versions of Maxwell's equations for light waves.

Quantum fields, of which light is perhaps the simplest example, are now regarded as being the fundamental entities of the universe, underpinning all types of material and non-material things. The only explanation is that the stuff of the world is neither particle nor wave but both. This is the nature of reality.

Chapter 5 Light matters

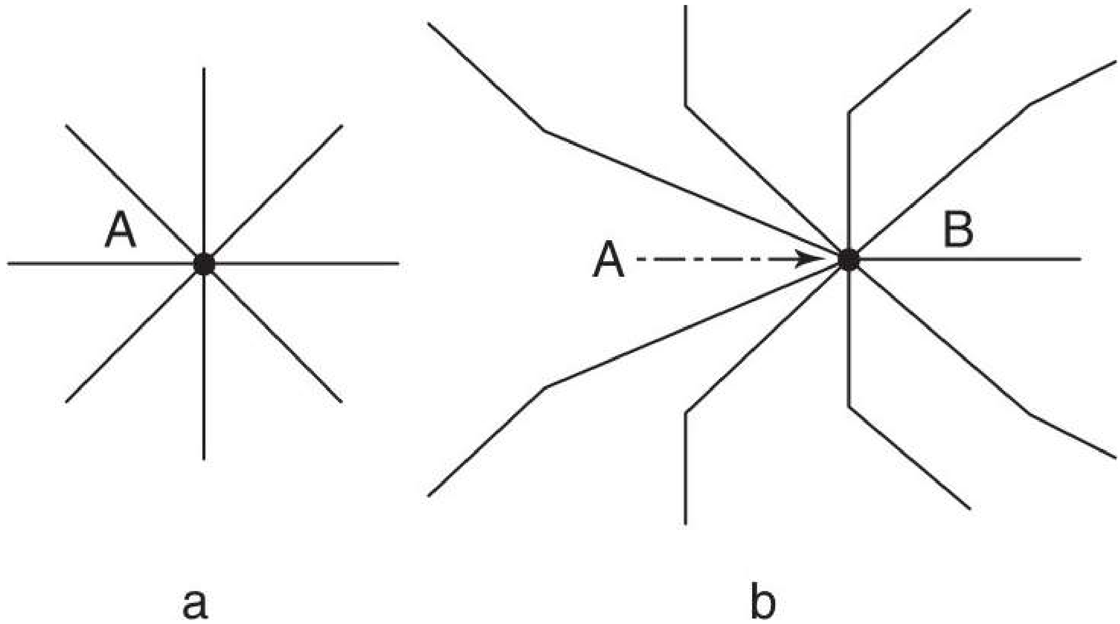
How do you generate light? There many different ways: common light bulbs, based on glowing metal filaments, as well as ‘fluorescent’ tubes; laser pointers; tell-tale lights on electrical equipment, from toasters to car dashboards; sunlight and starlight; even, for those in the northern and southern extremities of the planet, the aurorae; fireflies and glow-worms, and the phosphorescence in the wake of boats. What is the means by which these very different things all generate the same thing: light?

The key is that they all involve matter. More specifically, they involve electric charges moving about. When these charges accelerate—that is, when they change their speed or their direction of motion—then a simple law of physics is that they emit light. Understanding this was one of the great achievements of the theory of electromagnetism. An electric field has its origin in an electric charge, such as an electron in an atom. The electric field, attracting oppositely charged particles, such as protons, extends throughout all space, although it gets weaker quite quickly as you move away from the electron. As I noted in Chapter 3, this is the force that arises from static electricity.

Oscillating atoms and bending electrons

Now say the electron moves with a sudden jerk. The field surrounding it must move also, since it is linked inextricably to the electron. The change in the field is illustrated in Figure 25, as a ‘kink’. A remote proton will notice that the electron has moved only when the change in the field that arises from the electron movement has had time to make its way to the proton. This will take time since the kink, and therefore the information that the electron has moved, propagates at about the speed of light. When the change does get to the proton, it will move according to whether the electron moved closer to it (thus making the field experienced by the proton stronger, so that it feels a stronger force) or further away (leading to a weaker field and a weaker force).

Now say the electron were to move back and forth. Then changes would occur in the field surrounding it in synchrony with this oscillation and propagate away to the proton, which would then be induced to oscillate in turn as it received this information. But an oscillating electric field (and associated magnetic field, but that need not bother us here) is exactly what we mean by light.



25. Lines in the electric field of a. a stationary electron and b. an accelerating electron. When the electron accelerates, the changes in the field—the 'kinks' in the field lines—propagate away from the electron at the speed of light.

Since a hydrogen atom—the simplest one—consists of a single electron and a single proton, we can understand from this picture how atoms can generate light. First, let's think about what happens when light irradiates an atom sitting quietly on its own. The light forces both charged particles—the electron and the proton—inside the atom to move. But since the electron is much lighter than the proton, it moves more easily with a given application of force, and we can consider its motion with respect to the more or less stationary proton. In fact, the electron oscillates at the frequency of the light's electric field, being alternately accelerated or decelerated as the electric field varies.

This process is a bit like pushing a child on a swing. The best way to get the swing to oscillate is to push on it in synchrony with the natural oscillation period—a gentle push each time the swing reaches the bottom of its motion. Even so, it takes effort to get the child up to a height that thrills him or her. The child experiences maximum acceleration at the apices of the swing's oscillation, and maximum speed as she passes under the point of suspension. So it is with the atomic electrons—the energy of the light beam is absorbed by the atom, and converted into the motion of electrons.

Let's say you now stop pushing on the swing. What happens? The child gradually swings in arcs of lesser and lesser amplitude to rest. Again, so it is with the atom. The electrons gradually stop oscillating, and give up their motional energy by reradiating light. This is the process of light emission, and is the basis of many light sources, such as neon signs, fluorescent lights, and laser pointers.

Now in this picture I have assumed that some light beam itself is used initially to get the atomic electrons oscillating. But in a sense that begs the question of how one generates light in the first place. In fact it is possible to use other means to 'excite' the atoms. First, one can simply heat up the material. This is what is done in an ordinary light bulb, where by passing an electric current through a metal filament the metal heats up to a very high temperature—several thousand degrees. As the material heats

up, the electrons start to jostle and collide with the atoms and with one another more and more, and this both excites the atoms and causes the electrons to accelerate and decelerate rapidly. This produces a very broad range of colours that depends only on the temperature to which the material is heated, and not upon the particular type of atoms of which it is made.

Electricity may also be used in other ways. In light-emitting diodes (LEDs) for example, as used in displays, an electrical current—or flow of electrons through the material—can be captured directly by an atom, so the light is generated much more efficiently than in a thermal source. A fluorescent tube also uses an electrical current to directly excite atoms, only this time in a nebulous gas with which the tube is filled. Finally, many different chemical or biological reactions can release energy, some of which leaves the atoms or molecules in the form of light. This is the origin of the light given off by fireflies, for example.

We noted previously that acceleration has two parts to it: it can mean a change of speed, as we have just seen for the electron and proton in a hydrogen atom, or it can mean a change of direction without a change in speed. This latter characteristic of acceleration is familiar from the experience of going around corners in a car: you are pushed against the door or the side of the seat and experience a force that turns you with the car. The force is larger the faster you are moving as you enter the turn, and this is an

indication that you are accelerating, even though you may not be speeding up or slowing down.

Even this kind of acceleration, when experienced by charged particles, causes them to radiate light. Imagine a bunch of electrons forced to move in a circle, as if they were stuck on the rim of a rotating wheel. They generate light, the wavelength of which gets shorter and shorter (so the photon energy gets larger and larger) the faster they move around the circuit, because of this angular acceleration. The light generated in this way is called synchrotron radiation, and it is a common means to generate X-rays. It is also related to the light seen in the northern and southern aurorae, which are produced when charged particles from the Sun are forced to spiral by the Earth's magnetic field as they enter the atmosphere.

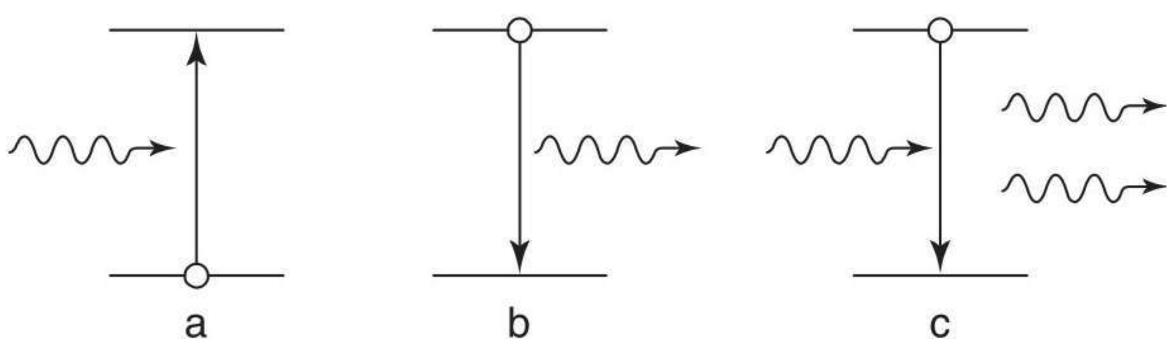
Quantum light emission processes

These basic mechanisms underpin all light sources. But the details of how the atoms act as a group can have a strong influence on the characteristics of the light that is eventually emitted—as I noted in Chapter 1, a light bulb emits a very different sort of light than a laser pointer, for instance. In order to understand this, we need to delve a little deeper into the structure of atoms, since the process of emitting light from an atom is not completely encompassed by the analogy with swings. Something needs to be added in

order to accommodate the fact that atoms and molecules are quantum mechanical entities.

For our purpose this simply means that the electrons in the atom or molecule can only hold energy in fixed amounts. Using our swing model, it means that the maximum extension of the swing amplitude cannot be anything you like. Rather, it is restricted to certain particular values: it is quantized. More particularly, the energy of the swing comes in discrete packets, or quanta, and when you push it you can only make it jump by one or more quanta. In the atom, this means the energy of the electron can change only in the same discrete units when it absorbs or emits a single photon.

The energies involved in making these jumps are very small by everyday standards. Take a light in your house. This consumes energy at a rate of, say, 60 W, or 60 Joules per second. A single photon emitted by the atoms in the light bulb possesses about one billion billionth of a Joule. So a light bulb is emitting more than ten billion billion photons per second.



26. An atom undergoing a. absorption, b. spontaneous emission, and c. stimulated emission.

An atom can be prepared in an excited state simply by shining light of the right frequency on it, as shown in Figure 26a. (Of course, this begs the question, but we’ve also seen that the excitation can also be accomplished in other ways, for instance by running an electrical current through the medium.) Now, according to quantum theory a corollary of the ‘jumpiness’ of the electron motion is that the atomic electrons are pretty much stable against the emission of light once in one of these discrete states. They are like a ball on a shelf in a cupboard—in principle it can lower its energy by dropping to a lower shelf, but this cannot happen in practice unless you give it a little push so that it rolls off the shelf.

So it would appear that quantum physics suggests that atoms cannot emit light, since once you have put them in these stable states—that’s it. Well, it turns out that, except for the lowestenergy state of the electron in the atom, a push is available to allow the electron to drop from a more energetic state to a less energetic state. And the surprising thing is that the shove arises from nothing.

In Chapter 4, I noted one of the strangest features of quantum physics: even the emptiness of space is a seething background of activity, filled with ‘vacuum fluctuations’. These fluctuations in the electromagnetic field can cause the atomic electron to drop to a lower energy level and to give

up the energy difference in the form of emitted light. This process, by which an atom (or molecule) undergoes a transition from one stable state (the excited state) to another of lower energy (the ground state), emitting a photon on the way, is called spontaneous emission (Figure 26b). It is something that each atom does on its own. It was originally proposed by Einstein in order to account for the proper energy balance between a light beam and the matter on which it was shining. If spontaneous emission does not occur, then the atoms hold on to the energy from the light beam and the situation we see everywhere around us, where most things are in a stable state in equilibrium with their surroundings, would be impossible.

Einstein understood the central mystery of spontaneous emission: that it was a random process. You just cannot say exactly when any given atom will make a jump. All you can say is that on average, after some amount of time (that depends on the particular atoms, but is roughly 1,000 billionth of a second), in a large collection of atoms, about two-thirds of them will have emitted a photon. But the origin of the fundamental randomness remained a mystery until, in 1927, Paul Dirac's quantum field theory introduced the idea that quantum vacuum fluctuations were at the root of this. The notion that a field containing no photons at all can cause an excited atom to be unstable is at odds with our intuition, and it took until the 1950s for Lamb's measurements to show that Dirac's explanation was right.

What this means is that even everyday occurrences—the picture generated on a TV screen by means of LEDs, for example—have at their heart this fundamentally random characteristic arising from quantum mechanics. By contrast, the emission of light by atoms when they are pushed to give up their energy by the application of another light field is called stimulated emission (Figure 26b). This form of recouping energy from atoms into the light field does not have random character. And this makes possible a very different kind of light: that of the laser.

Coherence: acting all together

When atoms and charged particles behave individually, ‘doing their own thing’, the light they emit when there are many of them is a sort of uncoordinated set of waves. Even a very small LED, with a size of much less than a millimetre, contains a vast number of atoms, so this is a common situation.

A feature of this uncoordinated emission is that each atom emits its photons at random, with no acknowledgement of what the adjacent atoms are doing: the light goes off in many different directions, and the photons are all emitted at different times. In effect, the randomness of the emission process is reflected in the randomness of the resulting light intensity.

Let’s say we put a photo detector in front of a light bulb. (A photo detector works like a light bulb in reverse.

It uses the photoelectric effect—light incident on it produces an electrical current that can be measured.) What we would see is that the electrical current from the detector was very noisy, because the light incident on the detector has an intensity that changes quickly and randomly, corresponding to the arrival of random numbers of photons at each instant of time.

But what if it were possible to coordinate the atoms, so that they acted together? We can think back to our earlier analogy: imagine a collection of swings, each with the same oscillation frequency.

The swings may be all oscillating at random: that is, with each at a different point in its repetitive trajectory at any instant. Or, they may be in synchrony, with the differences in trajectory between adjacent swings being a fixed amount—like a wave caused by the adjacent spectators in the crowd at a football match standing up and sitting down in sequence. In the first case, the light that is emitted from these uncorrelated atoms is like that from a light bulb or LED and is said to be incoherent. In the second case, however, the atoms oscillate in lock-step, and the light they emit is given off in a coherent fashion—all of the photons are emitted in the same direction. This is what happens in stimulated emission (Figure 26c), and is the basis for the laser.

Laser light

The laser is perhaps the most important invention in optics in the last century. This device produces extraordinarily useful beams of light that have revolutionized the range and capabilities of applications. Not only is it a specific source of illumination, in microscopy and spectroscopy for example, but it also provides a means to direct significant energy on to a particular target in a tailored way, and thereby to control the dynamics of matter. An extreme example of the latter application is laser-driven fusion of atoms, discussed in Chapter 7, that may enable new forms of nuclear energy which can draw upon a very large supply of fuel.A laser consists of an optical amplifier, or gain medium, that generates light from atoms by means of stimulated emission, placed between two mirrors (and possibly other optical elements) that form an optical cavity. The number of photons in the cavity builds up as the atoms emit light until there is a balance between the energy put into the light beam from the atoms, and the energy leaking out of the cavity through the mirrors. As the amplifier is turned on, by providing a means to excite the atoms, the light emitted from the amplifier is reflected back into it by the end mirror of the cavity. That causes further stimulation of radiation from the excited atoms, and thus the light in the cavity increases in brightness. At the other mirror, some of the light is transmitted out of the cavity as a useful output. Some is reflected back into the gain medium. When the rate at which light is put into the cavity by the amplifier equals the rate at which it is extracted through the output

mirror, the laser is said to be at threshold. Beyond this point, any increase in the amplifier gain (the rate at which atoms are put into their excited states) leads to an increase in the intra-cavity intensity, and thus to an increase in the output light.

The optical cavity also imposes a restriction on the colours of the laser. It turns out that the frequencies that experience the most gain are those for which the light waves add with constructive interference on each round trip. This means that the length of one round trip of the cavity should be equal to a multiple of half of a wavelength. The frequencies that satisfy this resonant condition are said to be the cavity modes.

The reason that lasers are important is related to the fact that the light they emit is coherent: all the photons go in more or less the same direction, with the same colour. The direction is defined by the laser cavity: the colour by the atoms in the gain medium and the allowed modes of the cavity. This leads to the property that the light is in the form of a beam—the laser beam—which is about as close to a ‘ray’ of light as you can imagine. It diverges as it propagates due to diffraction, but has the smallest divergence possible. This property also means that it can be focused to a very small spot using lenses or mirrors.

A second feature that contrasts the light emitted from lasers with that from light bulbs is that laser light is

usually very pure in colour. In other words, it consists of only a few wavelengths, whereas lamps often emit a broad range of wavelengths. The light intensity is very stable (registering as very low noise in a photodetector output), and the light can be emitted continuously or as a sequence of pulses.

The ability of laser light to be focused to a very small spot makes it useful for microscopy, and there are a number of different ways in which, by scanning the laser spot across an object at the focus of the microscope lens, and detecting the light scattered or re-emitted from the object, three-dimensional images of the object can be made. This approach is very useful in looking at animal tissue, for example, and optical microscopies of this kind are widely used in biomedicine.

Many applications of lasers in manufacturing also stem from this property of laser light. The ability to mark, cut, drill, or weld metals, for instance, requires that a lot of energy be deposited in a small region of the metal in a short time. So high-power lasers producing coherent light beams in the form of pulses that can be focused are ideal for such materials processing.

Similar properties are needed for some medical applications of lasers, also involving materials—this time skin, teeth, or hair. Laser correction of vision and laser dentistry are now commonplace, as is the use of lasers to

remove tattoos by heating up the ink until the drops break up, and for hair removal—although unfortunately not hair regrowth! Other familiar devices that make use of laser light's ability to address a very small spot are the CD, DVD, Blu-ray, and some computer disk storage devices. Very tiny spots of light in the recording medium allow the very high-density storage of data in the material.

The very pure colours achievable with laser light make it possible to distinguish the constituent atoms and molecules of different mixtures by means of spectroscopy. As noted in Chapter 1, different atoms, and indeed different molecules, have characteristic frequencies at which they absorb and radiate light, due to their different structures. Extending the analogy developed in this chapter, they are like swings in which the chains or ropes holding the seat are of different lengths—their natural oscillation frequencies are dependent upon the way in which they are put together.

In fact, each atom and molecule has a range of different absorption and emission frequencies, corresponding to the excitation of different electron configurations. Typically these lie in the blue part of the visible light spectrum, but some molecules absorb at much shorter wavelengths, invisible to humans. Many molecules also absorb light at wavelengths longer than the red end of the visible spectrum. This absorption arises from the vibrations between the atomic nuclei that make up the molecule. Since nuclei are much heavier than electrons, they tend to oscillate at much lower

frequencies. This set of frequencies is a sort of molecular 'fingerprint' that enables identification of a particular species.

The catalogue of these fingerprints is of importance, of course, in chemistry, since it allows the different elements involved in a reaction to be identified. It is also used in molecular biology, and even in cell biology, when the movement of particular 'tag' molecules can be studied. It is critical, too, for astrophysics, in which the elements present in astronomical objects—stars, galaxies, nebulae—can be determined, as well as in atmospheric physics and meteorology, for the remote sensing of pollutants or particles. Such monitoring provides key data in assessing the impact and origin of climate change.

By combining the outputs of several different lasers—say one emitting red light, one green, and one blue—it is possible to make a laser projector. By changing the intensities of each laser individually according to the video signal output of a computer or Internet link, perhaps by means of a liquid crystal cell, movies can be projected on a big screen with vivid, highly saturated colours. The combination of red, green, and blue (RGB) light is sufficient to make up a complete colour palette and lasers produce very bright images on a screen.

X-rays

When the wavelength of the light is very short, in the X-ray region of the spectrum, a different kind of spectroscopy arises. X-ray photons are energetic enough to excite the most tightly bound of the electrons in atoms—not just the outermost electrons. This means that X-rays can be used to look into the heart of atoms and molecules, and to understand their local environment, which can shift the binding energies of the electrons. X-ray absorption spectroscopy is widely used in the study of materials for a variety of applications, from detecting trace pollutants to understanding the structure of glasses. As noted in Chapter 3, X-rays are also used to study the structure of crystals by means of diffraction. When the X-ray wavelength is close to the spacing between atoms in the crystal, then the crystal acts as a 'diffraction grating' and scatters the X-rays in discrete directions. By detecting these diffraction patterns on a camera, it is possible to reconstruct the three-dimensional structure of highly complex crystals using advanced inversion algorithms. Today this is a routine process for characterizing isolated biologically and chemically relevant molecules, determining the structures of possible new molecules in order to design them for a specific function.

Some of the best light sources for this sort of spectroscopy are synchrotrons. In order to produce X-rays at the required short wavelengths, synchrotrons have to produce very energetic electron beams, and accelerate them around a big ring. Experimental stations catch a glint of light as the

electrons rush by, leading to short bursts of X-rays that are used for diffraction imaging. For example, the Diamond Light Source at Harwell in England accelerates electrons to more than a billion volts in a ring that is more than half a kilometre long. The next generation of X-ray light sources is being built using linear particle accelerators, which produce extremely bright X-rays beams. The X-ray diffraction pattern shown in Figure 20 was taken using the Diamond synchrotron.

Ultrashort light pulses

Laser light can also come in short bursts. There are several ways to arrange for this to happen. The method that produces the briefest pulses is called mode-locking. This requires a gain medium that has a large bandwidth—that is, it can amplify light over a broad spectrum. This allows several of the modes of the optical cavity to experience gain. If it is also arranged that these modes all have the same phase, then light waves of many different frequencies add constructively to yield a single pulse inside the cavity, bouncing back and forth between the mirrors. The brevity of the pulse is determined by how many frequencies are locked—the wider the range of frequencies, the shorter the pulse.

The possibility of creating very short duration laser light pulses enables a kind of measurement called dynamical or time-resolved spectroscopy. It allows us to see how things change in time, based on an old principle: the stroboscope. The general idea of how you use light to 'freeze' rapid

motion stems back to the work of Eadweard Muybridge in the late 19th century. He invented the idea of using a fast camera shutter to photograph a horse trotting.

The legs of a horse move too fast for the human eye to resolve, and it was not known at that time whether all four legs left the ground at any point during the stride. To settle the matter, Muybridge set up a bank of cameras, each of which had its shutter opened by a trip wire, which was triggered when the horse passed the camera. This enabled him to 'slice' a short pulse from the light reflected from the horse. The duration of this pulse of light was shorter than the time over which the horse's legs moved. The outcome of his work was that he was able to inform Leland Stanford, the funder of his research, that in fact there is a point in a horse's stride at which none of its legs touch the ground.

Mechanical shutters on conventional cameras could close very fast, but not fast enough to see some forms of animal motion, such as the flapping of the wings of a hummingbird. Even faster physical events such as an explosion, where things change on timescales of thousandths of a second, were out of reach. To solve this problem, Harold Edgerton at MIT in the 1950s invented a new kind of non-mechanical shutter, based on an optical switch. He was able to take 'still' photographs of explosions, for instance, using this device.

These shutters are what we might call 'passive' instruments. They simply allow a slice of light to pass

through when they are open, so are suitable to events that are well illuminated (the horse in California sunshine) or emit a lot of light themselves (the explosion). But we can imagine an ‘active’ instrument, one that generates a short pulse of light to illuminate a moving object. Think of the light pulses emitted by the flash unit on a camera. A flash of light with a duration that is short compared to the time taken for the thing we’re looking at to move provides an image of the object ‘frozen’ in time, even when using a camera shutter speed that is much longer than the changes of the object. A second flash freezes the motion at a later time. The same is the case for subsequent flashes.

A movie composed of the sequence of frames taken on repeated trials of the event reveals very rapid changes in the system, on a timescale that is much faster than could be observed by the eye. Indeed, the brevity of events that can be seen in this way is truly breathtaking. Edgerton invented a ‘stroboscope’ of this kind in 1931, and some of his most iconic images, such as a bullet passing through an apple or a playing card (Figure 27), were taken with this device.



27. A bullet frozen in motion using stroboscopic imaging.

Using modern pulsed lasers as the ‘flash’ it is possible to observe not just the frozen motion of bullets but even the motion of atoms in a molecule involved in a chemical reaction (for which the 1999 Nobel Prize in Chemistry was given to Ahmed Zewail) and the much, much faster motion of electrons whizzing around the nucleus of an atom. The timescales for these motions are staggeringly small—less than one tenth of a million-millionth of a second (100×10^{-15} seconds, or 100 femtoseconds (100 fs)) in the case of

molecules, and a few tens of billion-billionths of a second (100×10^{-18} seconds, or 100 attoseconds (100 as)) in the case of atomic electrons. These fields—femtochemistry and attoscience respectively—are at the forefront of what is possible in the interaction of light and matter. I shall consider them further in Chapter 7.

Chapter 6 Light, space, and time

Robert Grosseteste, Bishop of Lincoln and the ffrst Chancellor of the University of Oxford in the 13th century, was one of the leading thinkers of his day, and a proponent of the works of the ancient Greeks. For him, as for many philosophers, the challenge of understanding light's nature was critical to understanding the world. In Grosseteste's treatise on the subject, entitled De Luce, he extols the primary importance of light: 'The ffrst corporeal form . . . is in my opinion light. For light of its very nature difiuses itself in every direction in such a way that a point of light will produce instantaneously a sphere of light of any size whatsoever'.

For Grosseteste, light deffnes space by its propagation instantly throughout the universe. Without light, there is no space, and therefore no forum in which events can take place. Matter, and thus the spatial extension of objects, are coupled to light, but cannot be separately deffned. This intimate connection between light, space, and matter—in Grosseteste's hands amenable to quantiffable description—informed the development of ideas regarding cosmology in the subsequent centuries.

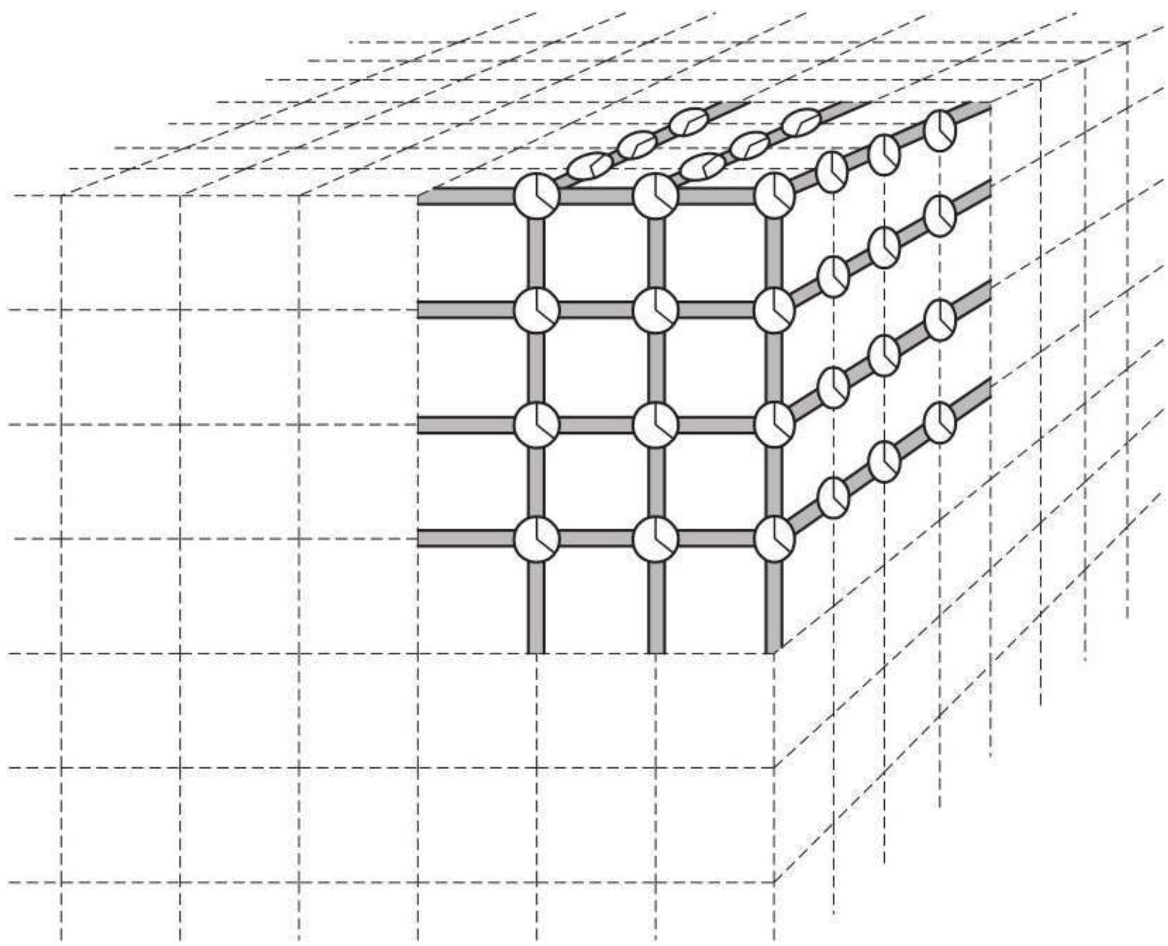
Space-time

For Newton, space neither admitted nor demanded deffnition. He thought of space as a pre-existing entity, a

sort of theatre in which events played themselves out. Large-scale motion of bodies in the heavens was integral to his idea of a set of universal laws. Einstein, by contrast, places light at the centre of space. For him, it deffnes space and time by virtue of setting the speed limit for signals sent from one part of the universe to another. The fact that there is a ffnite maximum speed turns out to make space and time inseparable. Einstein’s theory of relativity teaches us that we cannot think of one without thinking of the other. This is because our perception of space and time is based on local measurements of distances and time intervals. These measures appear difierently to those moving relative to us, because of the speed limit imposed by light.

How does this strange entwining of space and time by light arise? Let’s start with Newton’s conception of space. We can think of this as a sort of scafiold—a collection of imaginary rods of ffxed length all connected together in a three-dimensional framework, as shown in Figure 28. Newton thought that this sort of structure pre-existed any event, and indeed that all events took place somewhere in this structure. Events can therefore be speciffed by their distance from a ffxed point in the frame by counting the number of rods to reach the location of the event. Of course events occur at a certain time, too, so the scafiold must be equipped with clocks to measure the time. If a clock is placed at the junction of each of the rods, they all show the same time everywhere in space, and we can easily deffne a ‘universal time’.

Now we must ask several questions. First, how should we build a clock? Second, how should we ensure that they are all synchronized across space? Third, how should we build a ruler? These questions all have answers that are intimately related to the properties of light. Indeed, the answer to the last is this: one metre is the length of the path travelled by light in a vacuum during a time interval of 1/299,792,458 of a second. It is therefore linked to the answer to the ffrst question: how accurate a clock can be built.



28. A scaffold representing space made up of a 3D lattice of measuring rods. At each intersection is a clock, all synchronized.

Clocks

The primary characteristic of a clock is that it ‘ticks’ . That is, it signals at regular intervals of time. By counting the number of ticks between events, we can assign a time interval between them. The more precise a clock is, the more regular the intervals between the ticks. In a grandfather clock, the ‘ticks’ are provided by the regular swings of a long pendulum. In an electronic wristwatch, the ‘ticks’ are the oscillations of a piece of quartz crystal. These are much more regular than the swings of a pendulum, which can be afiected by the temperature and humidity of the place where the grandfather clock is located. Therefore the quartz watch is said to keep better time than the grandfather clock.

The most accurate clocks in the world are based on the highly regular clicks of electrons moving around in atoms. We’ ve seen that atomic electrons can jump between difierent stable energy levels within an atom. For some atoms, the difierence in energy between these levels is extremely well deffned, and the rate at which electrons jump between the two can therefore be used to deffne a set of ticks, based on the frequency of the light used to push the electrons up and down between the two energy states.

The idea is simple enough: illuminate an atom with microwaves (recall that microwaves are just like light, but with a much lower frequency—billions of oscillations of the

electric ffeld per second rather than millions of billions as in the case of visible light). Then slowly change the frequency of the microwaves until the electrons move between the atomic energy levels most eflciently. This defines the number of cycles per second of the microwave radiation—the rate of ticks of the clock—in terms of the energy level spacing of electrons in a particular atom.

There are many technical challenges to building such an atomic clock—cooling the atoms, preparing them in the right initial state, illuminating with microwaves in a clever way to maximize the sensitivity, detecting that the electrons are indeed in the higher state at a given time—but clocks based on caesium atoms are now the most accurate way to measure time, with a rate of ticking that deviates from complete and utter regularity by only one millionth of a second in one hundred million years.

These clocks provide a time standard that is agreed internationally and maintained by government laboratories such as the National Institute of Standards and Technology (NIST) in the US, the National Physical Laboratory (NPL) in the UK, and PhysikalischTechnische Bundesanstalt (PTB) in Germany. They are crucial components of many technologies that underpin our daily lives. For instance, they are vital for the global positioning system (GPS) that is the basis for navigation, including the satnav commonly used in cars.

Clock synchronization

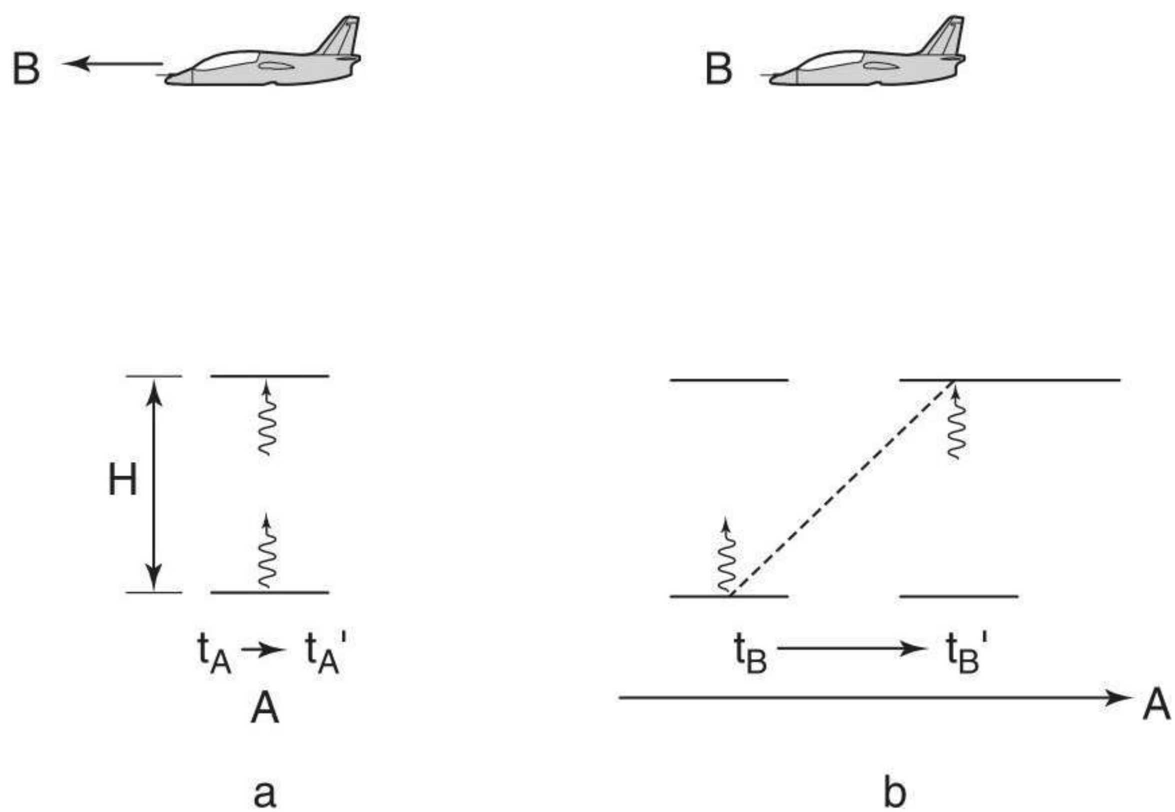
The next challenge is to synchronize the two clocks, so that they are calibrated. One way to do this is by sending a signal from one clock to another. You start the ffrst clock, sending a pulse of light to the other clock telling it when your ffrst 'tick' occurred. The person in charge of that clock then knows what fraction of tick her clock is behind yours (since the clocks are the same construction, we can assume they tick at the same rate) and can use this information in setting the correct time.

There' s an interesting consequence to this. Imagine you want to synchronize your clock on Earth to that on a distant planet, in a far-away galaxy. You send your pulse of light ofi into the direction of the planet, and then you wait. As the planet is far away, it might take a very long time for the light to get there, even given the high velocity of light. Meanwhile you are getting older and older. The person who receives your synchronization message will have received it from the young you—she will see you as you were when you sent the message.

Likewise when we look at the night sky, and see the distant stars, we are seeing images of them by receiving light that left their surfaces a long, long time ago. And as we look to stars or galaxies that are even farther away, so we look into the deeper and deeper past, seeing the universe as it was billions of years ago. In that sense, the light that reaches us is also billions of years old—it has been travelling across space for an age since its birth in the

remote past. Light is the oldest thing we can see in the universe.

Our clocks, though, are a bit closer together. An interesting fact is that when you put one of these clocks on an aeroplane moving at 800 kph or so, you ffnd that it ticks slower than one on the ground. That is, if you set the clocks to tell the same time, you will ffnd that the one in the aeroplane appears from the ground to be ticking at a slower rate than the ground-based one. This is a consequence of the maximum speed at which signals can be conveyed between the two clocks—the speed of light.



29. Time dilation due to relative motion. Observers A and B both measure the time for a pulse to arrive at the suspended mirror. They

measure different times because they are moving relative to one another.

You can see this by looking at the situation in Figure 29. There, person A is on the ground and person B is in an aeroplane moving at high speed. B watches as A sends a light pulse to a mirror suspended at height H above the ground. From A’s point of view, the light pulse travels distance H. However, from B’s perspective that signal will travel a slightly longer distance than H since A appears to be moving backwards at high speed relative to B. Therefore, because the signal travels at light speed according to both A and B, and both record the same number of ticks between sending and receiving the signal, then the only explanation is that B’s clock is ticking slower than A’s clock when seen from A, and A’s clock is ticking slower than B’s when seen from B. This efiect is called time dilation.

Einstein argued in a similar way that space should also contract. That is, a rod of B’s imaginary scafiold should look smaller to A than an identical rod in his possession. And vice versa: A’s rod looks shorter to B than her own.

Both of these efiects arise because there is a maximum speed at which signals of any kind can propagate, and this speed is the same for everyone. If that were not the case, then one could determine a preferred scafiold, or ‘frame of reference’, in which the signals went at the highest speed. Einstein’s work on relativity showed that there is no preferred frame, so that Newton’s idea of a ffxed, pre-

deffned space could not be the case. Since the maximum speed for signals of any kind turns out to be light speed, light is crucial in deffning space and time.

It's worth asking how precisely we might be able to synchronize two clocks by the 'light pulse' method discussed previously. You can see that one way to do it is by making the light pulse as brief as possible, so that the uncertainty over when it arrives is minimal. Thus it is important to know if there are limits to the brevity of light pulses. It turns out that there are, and they arise from a similar sort of wave property that limits the resolution of imaging systems, as we saw in Chapter 3.

We might start by asking how it is that we determine the frequency of a light wave. Imagine taking one of our clocks and asking how many peaks of the wave reach us during the interval between two ticks. The greater the number of peaks, the higher the frequency. The precision with which we can determine the frequency depends on how many times we repeat this measurement, since our ability actually to tell whether we have reached the peak of a wave may not be perfect. Thus, the longer we count, the more precise our determination of frequency. This trade-ofi is fundamental for waves—the imprecision in frequency multiplied by the uncertainty in the time interval is a ffxed product. This was ffrst understood by Joseph Fourier, a 19th-century French mathematician and scientist who played a key role in formulating the wave model of light propagation.

Ultrashort light pulses

Fourier’s theorem is important for clock synchronization, since it says that if we want a short pulse, we must have an indefinite frequency. Another way to say this is that a short burst of light is constituted by a broad range of colours. That’s entirely analogous to the situation Abbe identiffed for imaging optics: a high-resolution image, demanding a small focus, requires a wide range of ray angles to be collected. In fact, the analogy goes further. Just as Abbe showed the smallest size of a focal spot could be approximately one wavelength of the illuminating light, so Fourier showed that the shortest-duration pulse is a single cycle of the ffeld.

What this means in practice is that for light in the visible region of the spectrum, it is possible to produce pulses with a duration of about 2 fs. Amazingly, light sources derived from lasers can now routinely produce pulses of such startling brevity. They are based on the mode-locking technique described in Chapter 5.

But these are not the shortest bursts of light that occur in nature, nor even the shortest that can be produced in a laboratory. That honour goes to light sources with much shorter average wavelengths. For instance, using the single-cycle argument you can easily see that if the wavelength is shortened, then the duration of an optical cycle is reduced and in principle the duration of the pulse can be reduced.

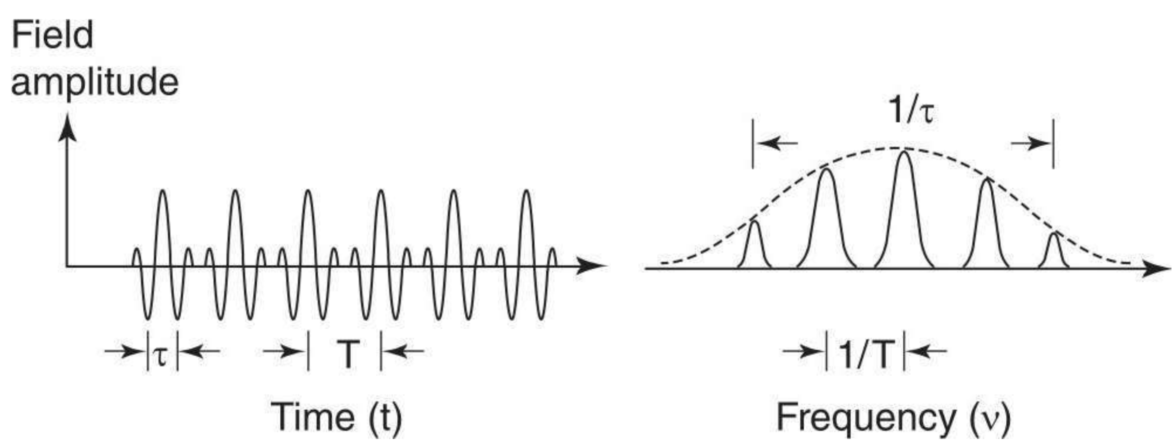
This approach currently holds the record for the world’s shortest controllable light pulses. By shining very intense laser light on an atomic gas, a process known as high-harmonic generation produces light waves with a frequency multiples of several tens of the driving laser frequency. This allows pulses with durations of several tens of attoseconds (10^{-18} s, or a billion billionth of a second). These are unimaginably short bursts of light, with duration equal to the time it takes an electron to oscillate within an atom.

Frequency combs

In Chapter 5, I stated that in a mode-locked laser a single pulse traverses the optical cavity. Each time it bounces off one of the mirrors, a little bit of the light pulse is transmitted through the mirror and exits the cavity. As a consequence, outside the cavity the light appears as a ‘continuous’ sequence of pulses spaced by the round-trip time of the light in the cavity, typically a billionth of a second or so. A ‘snapshot’ of such a train of pulses would show these very short bursts separated from each other by a delay that is long compared to their duration, like the teeth of a comb (as shown in Figure 30). And the pulses can be made identical to one another by careful adjustment of the laser producing them, so that the electric ffeld of each of the pulses peaks at exactly the same time with respect to the intensity envelope of the pulse.

It turns out that this conffguration means that each of the ‘teeth’ in the frequency comb has a very precise position at an absolute frequency. A precisely calibrated set of frequencies is a very important tool for building accurate clocks. This is because it allows a direct comparison of optical frequencies to lower (usually microwave) frequencies, which can be counted by means of electronics.

Thus frequencies in the microwave region inhabited by the caesium atomic clock can be compared simply to much more precise electronic transitions in the optical region of the spectrum in, for example, strontium atoms or aluminium ions. Therefore the standard caesium atomic clocks used in satellite navigation, for instance, can now all be synchronized and to tick at the same rate to within one part in a billion billion (i.e. 1:10ffi18), due to the precision of the optical electron oscillation frequency within strontium or aluminium.



30. Train of identical nearly single-cycle optical pulses. The spectrum of the pulse train looks like the teeth of a comb, hence it

is called a frequency comb.

‘Optical clockwork’ of this kind allows the comparison of disparate frequencies with such remarkable precision that it provides a means to test the tenets of relativity, and thus to understand better the role of light in deffning space and time. Frequency, and thus time, is the physical quantity that can be measured with the highest precision of any quantity, by far.

Optical telecommunications

Frequency combs are also important in telecommunications links based on light. In Chapter 3, I described how optical waves could be guided along a ffbre or in a glass ‘chip’. This phenomenon underpins the long-distance telecommunications infrastructure that connects people across difierent continents and powers the Internet. The reason it is so efiective is that light-based communications have much more capacity for carrying information than do electrical wires, or even microwave cellular networks. This makes possible massive data transmission, such as that needed to deliver video on demand over the Internet.

Many telecommunications companies ofier ‘ffbre optic broadband’ deals. A key feature of these packages is the high speed—up to 100 megabytes per second (MBps)—at which data may be received and transmitted. A byte is a number of bits, each of which is a 1 or a 0. Information is sent over ffbres as a sequence of ‘bits’, which are decoded by your

computer or mobile phone into intelligible video, audio, or text messages. In optical communications, the bits are represented by the intensity of the light beam—typically low intensity is a 0 and higher intensity a 1. The more of these that arrive per second, the faster the communication rate. The MBps speed of the package speciffes how rapidly we can transmit and receive information over that company’ s link.

Why is optics so good for communications? There are two reasons. First, light beams don’ t easily inuence each other, so that a single ffbre can support many light pulses (usually of difierent colours) simultaneously without the messages getting scrambled up. The reason for this is that the glass of which the ffbre is made does not absorb light (or only absorbs it in tiny amounts), and so does not heat up and disrupt other pulse trains.

Further, a light beam propagating in glass has to be very intense in order for it to inuence another light beam. For instance, when you cross the beams from two laser pointers you don’ t see either beam distort or deviate from its original path even though the two beams pass right through each other. (If your laser pointers had enormous power, you might just see such an efiect, but only because the room is full of air. In a vacuum they would still not inuence each other.) This means that the ‘crosstalk’ between light beams is very weak in most materials, so that many beams can be present at once without causing a degradation of the signal. This is very difierent from electrons moving down a copper

wire, which is the usual way in which local 'wired' communications links function. Electrons tend to heat up the wire, dissipating their energy. This makes the signals harder to receive, and thus the number of difierent signal channels has to be kept small enough to avoid this problem.

Second, light waves oscillate at very high frequencies, and this allows very short pulses to be generated, as described earlier. This means that the pulses can be spaced very close together in time, making the transmission of more bits of information per second possible. Indeed rates of 40 Gbps (a Gb, or Gigabit, is a billion bits) are possible in current-generation commercial long-haul systems. Electrical signals in copper wires are limited in the duration and spacing of pulses coding the information by the heating efiects noted previously, which tend to get worse at higher frequencies. Copper wires run out of steam, as it were, at much lower bit rates.

Fibre-based optical networks can also support a very wide range of colours of light. Glass transmits a broad range of wavelengths, with particularly low scatter and absorption loss in a spectral window from 1.3- 1.55 ffm. The rate at which photons are lost in ffbre at these wavelengths is about 5 per cent per kilometre. These losses can be made up by amplifying the light while in the ffbre so that transmission over very long distances (such as across the Atlantic Ocean) is possible without any conversion of the light to electrical signals or vice versa.

The telecommunications spectral window is divided into many individual frequency ‘slots’—much like the frequency comb shown in Figure 30. Each spectral component is a separate communications channel. There can be 150 or so slots in the window. Within each channel, a 40 Gbps optical signal can be operated. Therefore the total bit rate for communications is 150 times 40 Gbps, or 6 Tbps (1 Tb, 1 terabit, is 1,000 billion bits).

The ever-increasing demand for communications bandwidth due to increased use of the Internet and the services it provides have spurred optical engineers to new heights of creativity in harnessing the potential of light.

Chapter 7 Lighting the frontiers

Despite its long history as possibly the oldest continuous branch of natural philosophy and science, optics remains at the forefront of research and application. It is ubiquitous: as a tool for sensing, imaging, and communications, as well as providing ways to explore, discover, and illustrate new fundamental effects.

Light can generate conditions at the extremes of what is known to be possible according to physics, such as extremes of temperature and extremes of pressure and stress that do not exist naturally, except perhaps in the remotest of stars. And light can be used to observe and even control really fast events, such as the movement of electrons inside atoms.

Further, light can exhibit strange features associated with the quantum world, revealing even in everyday conditions some of the counter-intuitive aspects of the fitful world of randomness that underpins the stable, solid world of our normal experience. In this chapter, I will explore some of the frontiers to which, and across which, light is currently taking us. Exploration of these frontiers is possible because of the great technological strides that have been made in light sources, optical systems, and detectors, which enable exquisite control over the shape and intensity, in both space and time, a light beam can take.

Light mechanics

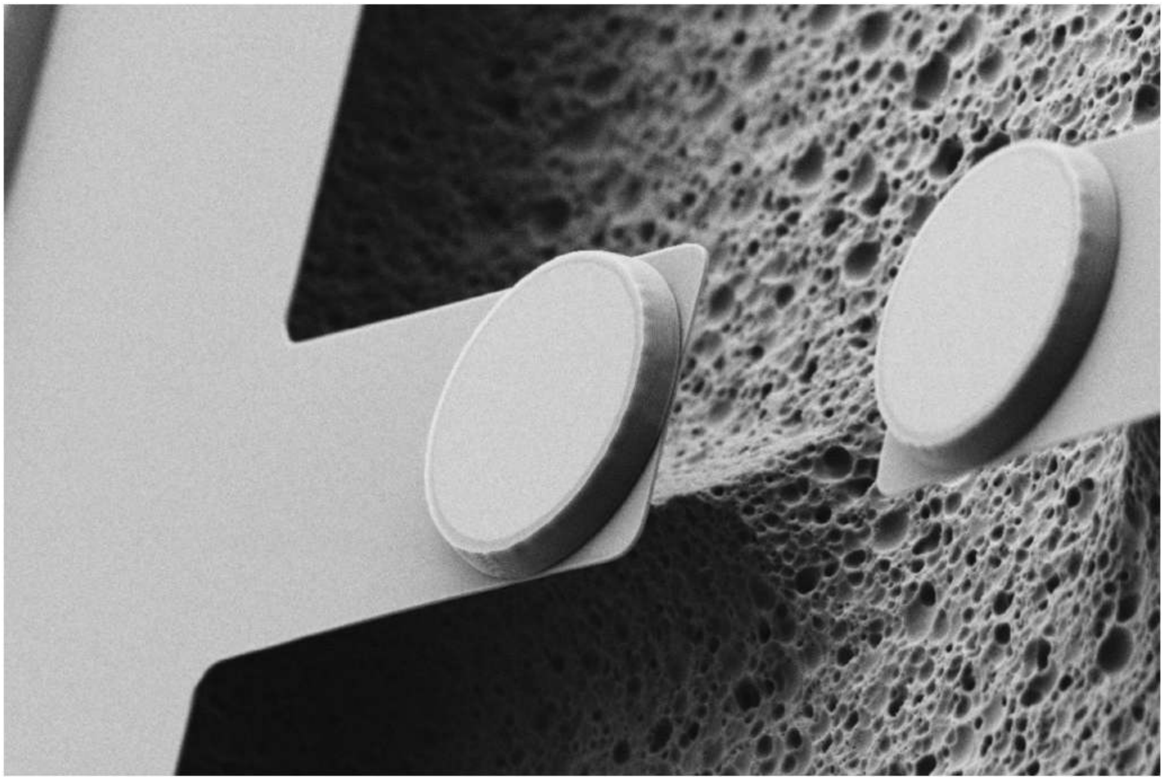
Light can exert forces on objects. This allows ‘remote control’ of bits of material using shaped light beams. Light can be used to move matter around, and bring it into contact with other objects, or to manipulate the internal configuration of atoms and molecules, forcing them into, for example, simple chemical reactions, in ways that allow both the study and exploitation of unusual material properties. That’s an extraordinarily powerful feature in many areas of research and study.

The concept of mechanical force arising from light has its origins in the momentum carried by each photon. For instance, when a photon is reflected from a mirror, that mirror experiences a force that provides the exertion needed to redirect the photon, just as water from a fire hose hitting a wall exerts a force on the wall by virtue of its bouncing off.

Similarly, when a photon is refracted it changes direction, and this, too, requires a force. Thus the photon exerts a force on the refracting element. If a light beam is incident on a glass bead, rays that are nearly tangential to the bead will experience the greatest change in direction. A photon traversing the ray in the lower half of the bead is directed upward as it propagates through the glass surfaces. The bead therefore experiences a force in the opposite direction. Since the momentum of the photon in the forward direction (the direction it was moving before encountering the bead) is reduced, there is also a net force in the

forward direction. The strength of this force depends on how many photons are refracted per second. A light beam that is more intense at a position near the centre of the beam than on its periphery will therefore drag the bead towards the higher-intensity part of the beam.

This effect can be used to make a focused light beam into an ‘optical tweezer’, which is able to hold on to a minute object and move it around as the light beam is steered. Optical tweezers find applications, for example, in biology, by enabling control of the position and movement of individual strands of DNA and the characterization of tiny molecular motors. Specifically, DNA, proteins, and other biologically important molecules can be stuck on to these beads and can therefore be handled using optical tweezers. Their position can be controlled with a precision of much less than the wavelength of light. This allows extremely small forces to be measured—such as happens when biological cells adhere to surfaces or other cells—as well as to hold the cells in places as they are operated on using other lasers (so-called cellular surgery). Optical tweezers can be combined with various other test methods, such as light scattering from aerosols, or spectroscopy, to characterize particles that may be pollutants in the atmosphere.



31. A nano-scale cantilever controlled using light forces. The discs are tiny mirrors about 30 μm in diameter.

These opto-mechanical forces can also be used to access completely new states of motion of small objects. In particular, it is now possible to build tiny mechanical cantilevers, illustrated in Figure 31, and to both observe and control their motion using light. Light forces can be used to cool or heat the oscillations of the cantilever—like running down or winding up a watch spring—and eventually to bring it into the quietest state possible, where only quantum fluctuations of the motion disturb the complete stasis of the lever. Light forces can also be used to cool atoms—much smaller objects—and this reveals even more strange quantum states of matter.

Ultra-cold

What’s the coldest thing you’ve experienced? Colder than winter in Oxford (approximately 2° C), or in Ottawa (ff20° C), or the South Pole (ff50° C)? Or perhaps the effects of liquid nitrogen, at ff200° C. These are certainly cold, but by no means the coldest things possible. It turns out that there is a lower limit for temperature: ff273° C, or 0 K (Kelvin), below which it is not possible to cool things further. This is the temperature at which things are as still as they’re going to get, with just the effects of quantum mechanics to cause a little jiggling about of atoms and molecules.

It’s not actually possible to build a machine to get to absolute zero but it is possible to get very close using an ‘optical refrigerator’. In fact, you can get cold enough to make the atoms almost stop moving. What this means is that their size gets bigger. (Quantum mechanics tells us that you can’t simultaneously specify the precise location and speed of an object. If the atom is completely stopped, then it must be extended over all space.) Therefore all atoms in the cloud that has been refrigerated occupy the same region of space, and this gives rise to some very strange new phenomena.

An optical refrigerator works by using lasers to ‘cool’ atoms. Imagine a laser beam shining on an atom that is moving from left to right, say. The laser shines from the right to left, so that a stream of photons hits the atom. The laser is

tuned in frequency to be absorbed by atoms that are moving at a particular velocity. Now, when the atom absorbs a photon from the laser beam, it gets a kick from the photon, and thus slows down. (More specifically, the momentum of the photon is transferred to the atom. Since it is in the opposite direction to the initial momentum of the atom, it reduces the momentum, and thus the speed of the atom.) The atom must re-emit the photon at some later time, and it will get a kick in the opposite direction to that in which it emits the photon. But the direction in which it re-emits the photon is random—it can go in any direction at all.

If you look at enough of these absorption-scattering events, then you will find that, although the light is always absorbed from one direction (the incoming laser beam) it is emitted uniformly in all directions—no one direction is preferred. The consequence of this is that on average a collection of atoms moving in a direction opposite to the laser beam grinds to a halt, and is left with random motion representing a temperature that is proportional to how long it holds on to the light before re-emitting it.There are several refinements of this approach, each of which uses light to cool atoms (and molecules) to even lower temperatures, and for which light acts like a 'viscous fluid' in which the atoms move slower and slower. It is even possible to use light to trap atoms using optical tweezers once they are slow enough. This allows the application of yet more sophisticated optical cooling techniques, by which it is

possible to get to temperatures of a billionth of a degree above absolute zero.

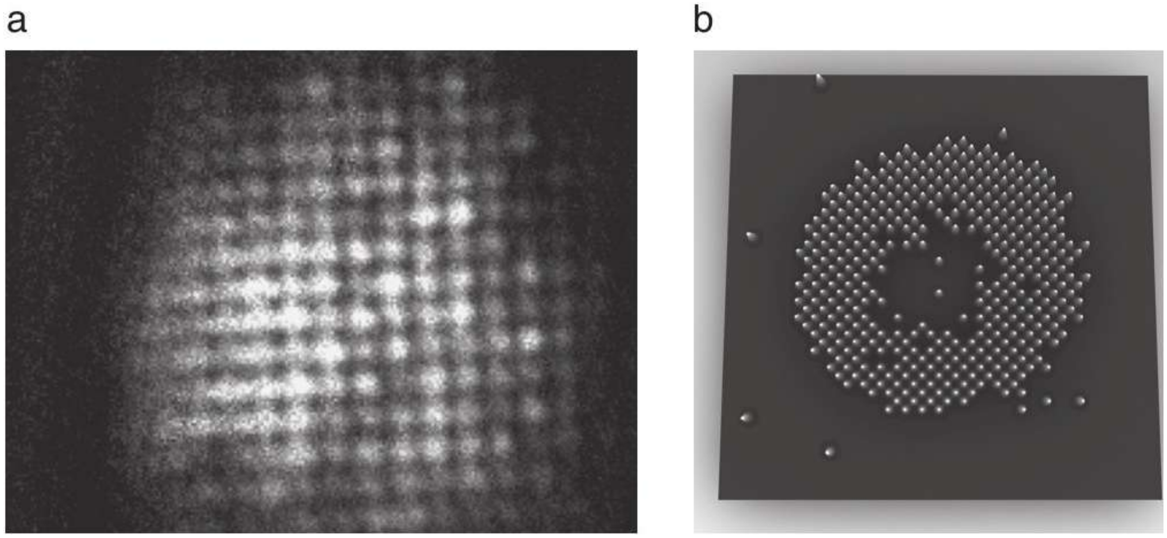
I referred to some residual ‘jiggling’ of the atoms that happens even at zero temperature, arising from quantum mechanics. The range of this jiggling can be thought of as the spatial extent of the atom itself. That is, according to quantum mechanics the atom isn’t just wandering around in a random fashion over a small region of space, but rather it is actually present across all that region at once. For atoms trapped at such low temperatures the size of that region may be several thousandths of a metre. That’s a remarkably large atom, given that the distance of the electron from the atomic nucleus is less than one tenth of a billionth of a metre. What’s even stranger is that several atoms can occupy this region of space at the same time.

That’s conceptually very counter-intuitive. We often think of atoms as being like little billiard balls, that can be packed close together, as in a solid material, but which retain their individual distinction by virtue of their location inside the material. That’s not the case for these very cold atoms. They are each everywhere at once, in a new state of matter identified by Einstein and the Indian scientist Satyendra Nath Bose and called, not surprisingly, a Bose- Einstein condensate.

This very strange state has some remarkable properties. For instance, it is a superfluid, which flows without

viscosity. Further, it is possible to split the entire atomic cloud in half and recombine it to show quantum interference between the two separated parts, essentially demonstrating that a big object (containing many atoms and of a palpably visible size) exhibits quantum character, attributable to the uncertainty of whether an atom is in one part of the cloud or the other. One has to think of each atom occupying both separate components at the same time.

Because these cold atoms can be trapped in light beams, it is also possible to create spatial structures out of several light beams that can be used to manipulate the atoms. For instance, when two light beams coincide they form an interference pattern (see Chapter 3) in which there are regions of high and low intensity. Cold atoms like to settle in one or other of these regions (you can adjust which one by choosing a particular wavelength of the light). As the intensity of the light beams is turned up, the atoms fall into the 'egg-crate'-like optical traps that appear in the intensity pattern, as shown in Figure 32a. And they do so in very interesting ways.



32. Cold atoms trapped in an optical lattice: a. a few hundred atoms per cell (several 10s ffK), b. individual atoms at each site (a few nK).

When the atoms are cold enough, they don’t like to be located at the same ‘site’ in this egg-crate, so the resulting distribution of atoms is very like a full egg carton—one atom at each site, as shown in Figure 32b. In this case there is no superfluidity, since the atoms like to stay put. It is more like an ‘insulator’, as nothing moves. By turning the light intensity up and down it is possible to explore this interesting transition between completely free flow and no flow at all.

The ability to do this in a system that is fully quantum mechanical allows scientists to explore new properties of matter that are relevant to other types of materials (for example, solid-state metal oxides) over which it is difficult to exert the same degree of precision control and

measurement. In cold atomic gases it is now possible to look at atoms in these egg-crates individually and see what they are doing as changes are made to their environment.

It is possible to explore this low-temperature regime with many different kinds of atoms, and to build complicated trapping structures using light. The idea of using cold atoms to ‘simulate’ other quantum systems is a current area of research. It allows exploration of complex problems that cannot be solved in other ways, and is expected to lead to a new understanding of materials and structures that will have impact in new ways—perhaps helping to understand and even design new magnets that will be used for applications such as data storage for computers, magnetic resonance imaging machines for healthcare, or even friction-free motors for levitating trains.

Ultrafast

Light pulses can be extremely brief. In Chapter 5, I stated that they can be as short as a single cycle of the optical field. For light in the visible region of the spectrum, that’s about 2 fs. For light in the extreme ultraviolet region, which is of shorter wavelength and higher frequency, the durations can be much shorter. The shortest yet measured is less than 100 as (1018 s) long. These are currently the shortest pulses that can be controllably generated (although we can observe events that happen on a much shorter timescale by means of particle colliders). And

with the advent of bursts of light in the X-ray region of the spectrum, we can expect that even shorter timescales are possible.

Since these numbers are so mind-bogglingly small, it is helpful to put them into context. The age of the universe is approximately 5 × 1017 s. Thus the ratio of one second to the age of the universe is approximately the same as the ratio of one attosecond to one second. Or to put it in an economic context, if the national debt of the US is equivalent to a second, then one cent would be equivalent to a femtosecond. On this scale, a single attosecond is virtually worthless.

What sort of things can happen on this timescale? In Chapter 4, I introduced a simple model of an atom—called the Bohr model—in which electrons ‘orbit’ an atomic nucleus, attracted to it by electric forces in much the same way as planets orbit the Sun, attracted by gravitational forces. The time taken to execute these orbits for simple atoms (that is, those with only a few electrons) is about 150 as. So if we want to look at this motion, we might need to use pulses shorter than this, so we don’ t just see a blur.

The idea of the stroboscope is the one most relevant to our story, since a variant is used by researchers today to look at the really speedy changes that go on at the fundamental level of atoms and molecules. In this application the light pulses from a laser are split into two (or more)

parts, with a delay introduced between them. The first pulse in the sequence illuminates the sample, and some of it is absorbed. This ‘triggers’ some changes in the system—electrons move around in the atom, or bonds vibrate in molecules or solids. An instant later the second pulse arrives and some of it is again scattered from the sample and detected.

As the delay between the two pulses increases in repeated runs of this experiment, the detected scattered light maps out the dynamical changes of the sample. In a sense it makes a ‘movie’ of the atom or molecule or solid as it changes. This ‘pump-and-probe’ scheme has been used to get into the guts of what happens, for instance, during a chemical reaction, when two molecules are reconfigured by their interaction. More sophisticated versions of this kind of approach exist using several light pulses. These approaches are being used now to explore many fundamental features of extremely interesting and puzzling materials, from interacting atoms and high-temperature superconductors to biological systems.

I’ve noted that the shortest pulse it is possible to generate is a single cycle of the optical field. It turns out that you can devise an experiment to measure the oscillations of the optical electric field using the extreme ultraviolet (EUV) pulses produced by highharmonic generation. What is needed to measure the pulse field is a very fast process, one that is much faster than the optical cycle itself. That can

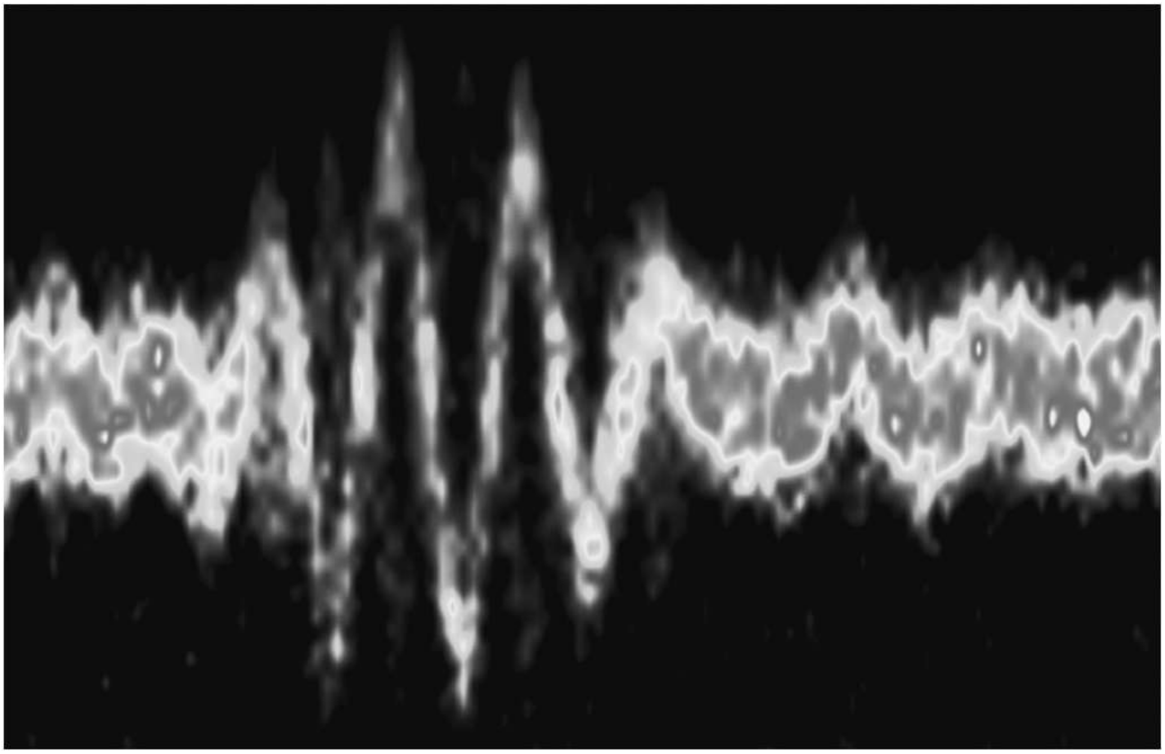
be provided by a pulse with much shorter wavelengths, about twenty or thirty times shorter than that of the optical wavelength. Pulses of such brevity are generated when an electron is ripped off an atom by means of a strong optical pulse. This requires an optical field whose strength is comparable to the binding force of the electron to the atomic nucleus. Such pulses are readily available by adding optical amplifiers to the output of a mode-locked laser.

When the electron is liberated by such an intense pulse, it finds itself sitting in a rapidly oscillating electric field, and, if its liberation occurs near a time when the field has zero amplitude, the electron can 'surf' along the next cycle of the optical wave, taking an excursion away from the atom, and then back again. When it returns it is moving very fast, and when it recollides with the atom it can be recaptured by emitting all of its extra energy as light. In this case a very short pulse is emitted as the electron recombines, having a very short wavelength of perhaps a few tens of billionths of a metre, in the EUV region of the spectrum, about twenty times shorter than the optical wave that generated it.

Now imagine that this EUV pulse shines on another atom. It has a sufficiently short wavelength such that it is absorbed by the atom, and knocks off an electron, which then sits near the atom. Imagine further that the atom is simultaneously illuminated with the short optical pulse we seek to measure. The field of this pulse accelerates the

electron in one direction or another depending on the part of the optical cycle at which the electron is liberated by the EUV pulse. By changing the delay between the EUV pulse and the optical pulse, the acceleration of the electron can be measured since faster electrons, which have been accelerated to a greater degree, have more energy. In this way it is possible to 'see' an optical pulse field (Figure 33), despite the extraordinarily short timescale of the oscillations of the field.

An example of an application of the methods of pump-and-probe spectroscopy in biochemistry is the study of the first steps in the process of photosynthesis, by which plants convert carbon dioxide from the air into oxygen using sunlight as an energy source. The processes by which this happens involve transporting energy around a big biological molecule with remarkably high efficiency. The means by which this happens has some very interesting and poorly understood features—it is faster than one might expect and much more efficient. If we could learn from systems that have evolved naturally over eons how to do this, perhaps our understanding would enable us to apply it to things like improving the design of solar cells, which would have an enormous impact on society.



33. Direct image of the electric ffeld of an optical pulse. The time between two adjacent peaks is approximately 2.5 fs.

Ultra-intense

Your electricity bill tells you how much energy you used in the previous month. It is measured in units of kilowatt-hours (kWh), and you are charged for each unit that you consume. Let’s say that you used 220 kWh in some month (this is the average monthly energy consumption in the UK). Now, you could use all this energy at the same rate across all four weeks of the month. Or you could use it in the first week, and use nothing in the subsequent three weeks. But can you imagine using it all in a million billionth of a second? You’d need to have an awful lot of appliances to use that

much energy in such a short time, and you'd need to be able to switch them on and off impossibly quickly. But the peak rate or power would be immense.

It is possible to produce light pulses that can achieve this. That is, they are incredibly brief and contain this amount of energy. In fact, it is possible to produce a pulse that delivers energy at a rate equal to the entire electricity generating capacity of the planet at a given instant. But the lights in your house won't go out, because the pulses are so brief that the total energy in them is very small.

Lasers that produce pulses of this kind are massive instruments, occupying large buildings that are a significant fraction of a football field in size. One example is the VULCAN laser at the Rutherford Appleton Laboratory in England. VULCAN produces pulses with 500 Joules of energy (1 J = kWh) in a pulse of 500 fs duration. 500 J is the energy emitted by a 100 W light bulb in five seconds. Yet the brevity of the pulse means that the intensity of the light can be as high as a million suns. The laser at the National Ignition Facility (NIF) in Livermore, California is much, much bigger than this. And the proposed European Light Infrastructure project is set to deliver a system capable of even greater peak power than that at NIF.

Such very brief, very intense light bursts can be used to alter states of matter. The electric field at the most

intense moment of the light pulse is larger than the attractive field between the electrons and nuclei that holds atoms together. So it is possible to strip the electrons off the atoms to form a new state of matter—a plasma. And it’s possible to do this in an instant, shorter than the time over which the atomic nuclei can move, so that the plasma is very dense—nearly the same density as in a solid block of material such as a piece of glass, except now at two million degrees Celsius.

These are the conditions inside the cores of the giant planets and even some stars: very high-density plasmas with particles colliding with each other at high speed and at pressures of a million times that of our own atmosphere. It is possible to use this new laboratory-accessible state of matter for several things. For instance, we can begin to understand how stars work, what their life cycle is, and to characterize their stages of evolution, such as supernovae explosions and white dwarfs. Other situations of interest to astrophysicists are also amenable to experimental exploration using lasers. Astrophysicists also use such plasmas to probe the frontiers of planetary science. For instance, it is possible to infer the composition of the gas giants from their mass and size, but only if the degree to which matter can be compressed at such high pressures is known.

Some laser facilities generate pulses using light of very short wavelengths. These pulses are produced by accelerating electrons in magnetic fields, so that they ‘wiggle’ from

side to side as they whiz down the accelerator. This produces a kind of synchrotron radiation consisting of short bursts of X-rays. Often these kinds of lasers use techniques, and indeed hardware, from particle accelerators. Examples of these are the Stanford Linear Collider Light Source (LCLS) and the Hamburg X-ray Free Electron Laser (XFEL).

The techniques afforded by the most intense laser pulses, in conjunction with short bursts of X-rays, enable scientists to diagnose plasmas in a wide range of conditions. Further, the immense pressure exerted by lasers between atomic nuclei can cause them, under the right conditions, to fuse together, releasing a large amount of energy in the process. This ‘nuclear fusion’ is a possible route to an almost unlimited source of energy. This application of lasers is extremely technically demanding, and is one of two methods that are being explored to achieve fusion: the other does not involve light except as a way to monitor the process. Both make use of dense plasmas.

As the laser pulses move through these plasmas, they generate a wave, very similar to the wake that trails a boat moving across a water surface. The electric fields in the plasma wave can reach more than 100,000 V over 106 m (that’s about ten times the voltage in the big power lines on pylons, over a distance of less than one tenth of a human hair’s breadth.) These field strengths are at least 1,000 times bigger than the accelerating fields used in the world’s biggest machines for studying fundamental particles, such as

the Large Hadron Collider (LHC) at CERN in Geneva. It may be possible using laser methods eventually to build table-top devices that can accelerate electrons to similar energies as currently possible in the LHC.

It is also possible to accelerate heavier particles, such as protons, by means of the extraordinarily strong electric fields generated by the interaction of laser pulses with matter. Proton beams are currently being explored as a cancer treatment, since the delivery of heavy particles to diseased tissue can be done more precisely and at greater depths than is currently possible using other types of radiation therapy.

The extraordinary properties of light continue to enable new realms of discovery, across a wide range of fields. Light is a ubiquitous tool for science and technology.

Chapter 8 Quantum light

In Chapter 1, I introduced the idea that light could be construed as a stream of particles, which I labelled ‘photons’ for convenience. It turns out that these are real particles, which can be produced, played with, measured, stored, and used for doing things. However, even though photons are in a sense the simplest expression of light, making individual photons is not so simple. Most light sources generate light of a different kind, for which the number of photons is not fixed.

A light bulb, for instance, produces a stream of photons that sprays everywhere. If you looked at the light going in just one direction from the bulb, and then examined just a short temporal section of the beam—a time slot, if you like—then you’ d be able to count some photons in that slot. But if you repeated the experiment several times, you’ d find that the number of photons was random,

sometimes large and other times small. The average number of photons would be fixed, depending on the brightness of the bulb, but you’ d never be able to say with certainty how many photons you would measure in the beam at a given time. That’ s one of the characteristics of ‘classical light’ — light that can be described entirely in terms of waves.

Laser light is also of this kind. The average number of photons in a pulse of laser light can be large, but for any

given pulse the actual number of photons will be bigger or smaller than the average. The spread of photon numbers in a pulse is approximately the square root of the average number, so that the relative ‘noise’—the variation in the number of photons in each pulse compared to the mean number over all pulses—gets smaller the higher the average number of photons.

Thus a laser beam has intrinsic intensity noise. This sets a limit on the quality of images you can get with laser illumination. The fluctuations in the laser intensity mean that detecting the separation of two points in an image is imprecise. In fact it is very imprecise for low-intensity light, where the mean photon number is small (so the object is hard to see) and the variation in photon number from frame to frame is large. The only way to get precise measurements is to look for longer, thus increasing the number of photons illuminating the object, and averaging the results over many laser pulses. The relative intensity noise is reduced by this signal averaging, leading to a better-resolved image. The precision increases in proportion to the square root of the number of photons used. This is called the ‘standard quantum limit’, since no classical light beam can beat it.

Quantum light, on the other hand, allows you to achieve much better results in signal averaging for the same average number of photons, since quantum light can have much lower noise than any classical light. But first, you have to build a quantum light source. There are many kinds of such a

source, each producing a distinct kind of quantum light. But we might consider, to be concrete, a source that generates the primitive quantum state of light—a photon.

Single photons

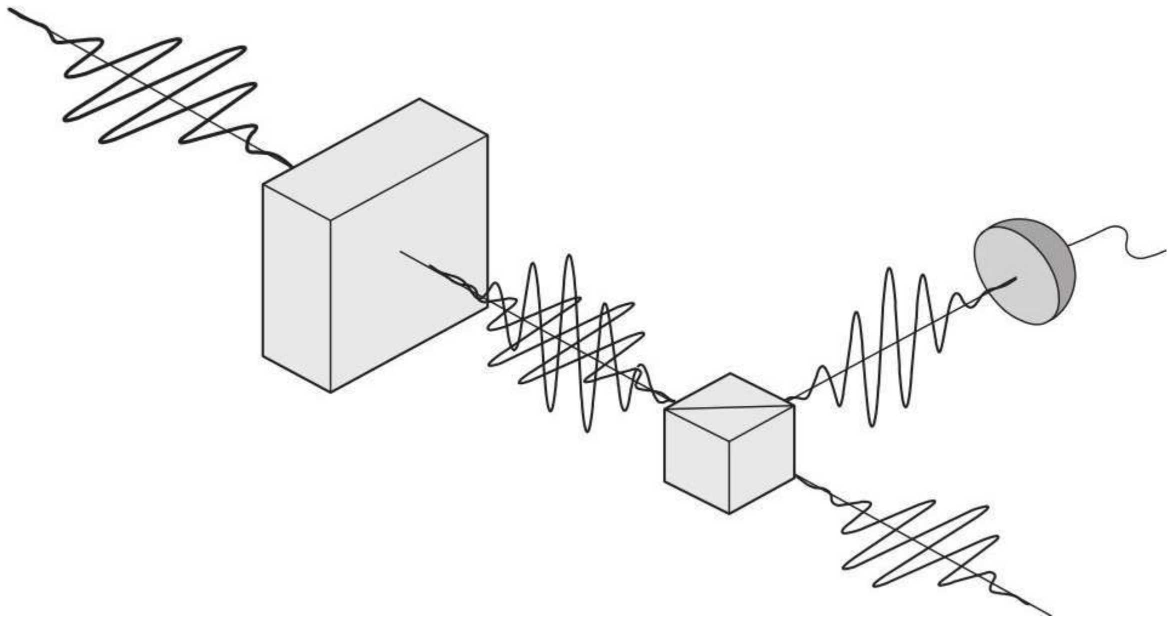
So how could you make just a single, individual photon? There's a very practical scheme, invented by Otto Frisch in 1965. His idea was simple. Take a single atom and put it in its excited state (see Chapter 5 for a discussion of how to do this). Then wait for it to drop to its ground state. When it does, it emits just one photon, since only one 'quantum' of energy can be stored in a single atom. You can tell when the atom has emitted the photon, because it recoils from the 'kick' provided by the photon's momentum. If you detect the atom moving, you can determine both that the single photon is on its way and the direction in which it is going.

Some modern quantum light sources operate in a similar way to this, only they corral the atom between two mirrors (an optical 'cavity' similar to that of a laser), and excite it very quickly so that it emits preferentially into a direction defined by the cavity axis. This makes a reliable source of single photons. It is an especially 'low noise' source, since the photons are emitted with strict regularity. If you looked at a given time slot in such a beam, you'd be able to predict with certainty how many photons would be in it—just one. Therefore the intensity is exceptionally stable

—it is a ‘quiet’ light beam, in contrast to the ‘noisy’ classical one.

The idea is also used in other quantum light sources. In particular, you can construct a very simple light source using nonlinear optical effects. Specifically, there are crystals that enable one input photon with high energy to be split into two photons with lower energy, each about half of the original input photon. The probability that this fission takes place is rather small, for most materials. But since the photons are produced in pairs, you can use one as a ‘herald’, to signal the presence of the other (Figure 34). Such light sources are the workhorse of the field of quantum optics, which uses the quantum mechanical features of light to explore the foundations of quantum physics, as well as to enable new kinds of information technologies.

Just as classical electromagnetic waves can be polarized, so can photons. So we might find a vertically polarized (V) photon or a horizontally polarized (H) photon. These would behave just like waves, in that if we measured the polarization of the photon by seeing if it passed through a polarizer oriented horizontally, then we’d find that the H-photon always passed through and the V-photon never.

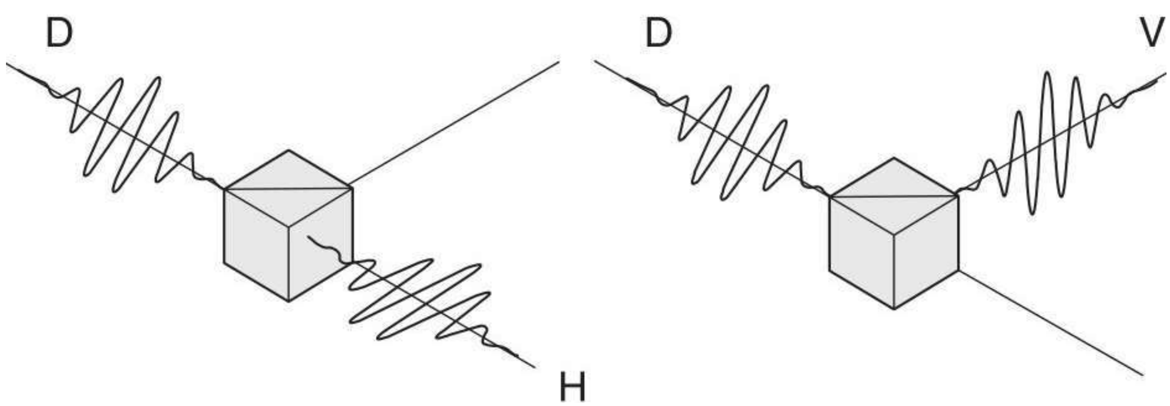


34. A ‘heralded’ single photon light source, generating photons randomly, but with a signal that indicates when one has been prepared.

What’s strange here is that we can construct a diagonally polarized photon, oscillating with the field at 45 degrees to both the horizontal and vertical. But if we now try to see if the photon passes through the horizontal polarizer, then there is an ambiguity. The photon is the smallest ‘piece’ of light, so can’t be divided further. How should it behave at the polarizer? What happens is that it is transmitted with a probability of one-half, and reflected with equal probability (illustrated in Figure 35).

In practice what that implies is that if you try the experiment of putting a diagonally polarized (D) single photon into a horizontally oriented polarizer a million times, then 500,000 times it will go through. And the very

strange thing about quantum mechanics is that you cannot tell on any given trial what will happen. This is not because the photon can be considered sometimes to be H-polarized and sometimes V-polarized. Rather it is because the photon is both H and V-polarized, simultaneously. The random outcomes of a measurement of the photon’s polarization therefore reveal the intrinsic uncertainty that inhabits the most fundamental level of the universe as described by quantum physics.



35. A diagonally polarized photon encounters a polarizer, and exits randomly through one output or the other.

Of course, you can make a virtue out of necessity in such circumstances. You can do practical things with single photons that are unimaginable with ordinary light. For instance, this property of photons can be used to generate random numbers, by measuring whether the photon is transmitted (labelling the outcome, say, 1) or reflected (labelled 0). The randomness in the string of zeroes and ones is inherent in the underlying physics, not just in the

manufacture and casting of dice, or other contingencies. For this reason quantum random number generators are an emerging business—you can’t fake the randomness they provide.

A second example: you can make communications links for which the security is guaranteed by the laws of physics, rather than by trusting your telecoms supplier. This is because of two important properties of photons. First, you cannot detect them in two places at once. For that reason, if an eavesdropper grabs the photon to capture the information you are sending, then of course you don’t get the photon. So you receive no information, and you are aware that something’s wrong.

But if the eavesdropper is clever, she will send a ‘decoy’ photon that she hopes will fake the message. But you can tell that it’s a fake! The reason you can know this is that in quantum mechanics there is no measurement that can tell you everything about a single quantum particle.

Consider the following scenario. You want to send a simple binary message (0s and 1s) over this link, say a vertically polarized photon for 0 and a diagonally polarized photon for 1. If the eavesdropper (usually known as Eve) measures the photon and gets the answer ‘vertically polarized’, she could not be certain that the photon was a 0, since the diagonally polarized photon would give her the same answer at least half the time. So she gets some information, but not everything.

Now, let’s say the sender of the message (conventionally called Alice—you, the receiver, are Bob) sends you a photon coded as 1. Let’s say Eve measures this in the vertical orientation and gets a positive result. She must choose whether to send you a vertically or diagonally polarized photon. One strategy is to send you a vertically polarized photon, since that’s the most likely source of her result. Now you measure the diagonal polarization. If your photon is from Eve, it will give you the wrong result 50 per cent of the time. If it is from Alice, you will never get the wrong result. So by comparing a section of the received message with what Alice sent, you can tell if Eve is tampering with your line.

However, Eve might be even cleverer. She may try to copy the photon from Alice without measuring it. She could make two copies in fact, sending you the original. Then she can make a vertical polarization measurement on one copy and a diagonal polarization measurement on the second, and she would have determined the full information about the photon ‘bit’ that Alice sent you without you ever knowing. However, she would be thwarted. A remarkable feature of quantum mechanics is that there is no possibility of building a copying machine that can do this—make an exact replica, or clone, of a single particle in an unknown quantum state. It is simply forbidden by the laws of physics. Because of these two constraints imposed by physics— ‘no measurement’ and ‘no cloning’—it is possible to build a secure

communications link that can transmit a secret stream of random bits between Alice and you.

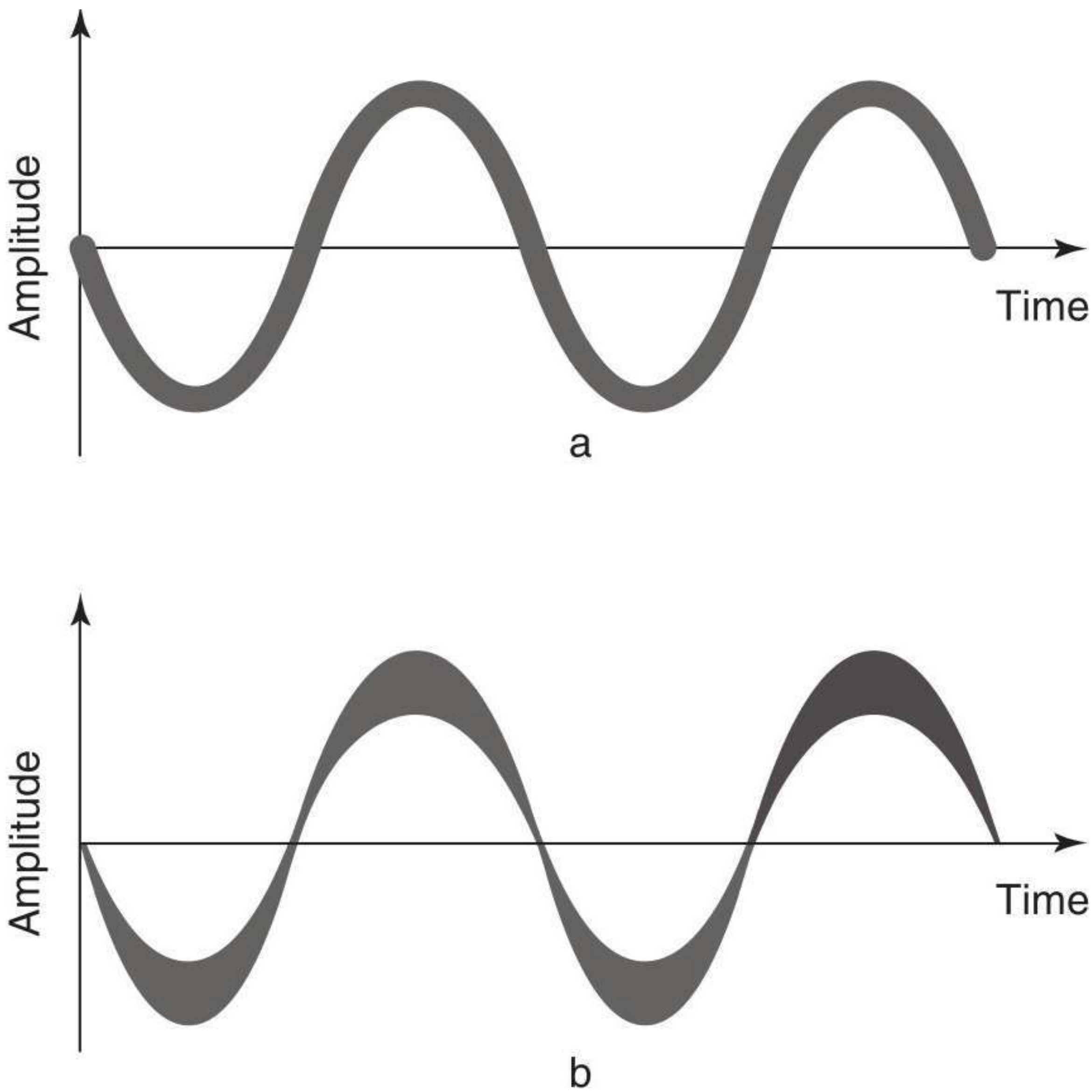
Squeezed light

There are other kinds of quantum light that provide different sorts of enhancements. Recall that light is oscillations of the electromagnetic field. A laser beam most nearly mimics this ideal behaviour. Yet even it has some ‘noise’ in the amplitude. That is, each time you measure the field amplitude, you get a different answer. The situation is sketched in Figure 36a, which shows some uncertainty about the field at each point, or phase, of its oscillation. There is a particular type of quantum light—called ‘squeezed light’—for which this noise varies with the point in the cycle of the field, as shown in Figure 36b. It is bigger at some phases than at others. It turns out that such a field is composed of only pairs of photons. If you measure the number of photons, you will only ever find an even number. The quantum interference of all these pairs is the origin of the phase-dependent amplitude noise.

There are certain things you can do with such a state. Imagine that you wanted to make a measurement of the phase of the wave. (Recall that that is what you do in an interferometer, and the phase shift is something induced on the light beam by an object you’d like to measure, such as the presence of a particular molecule.) The phase can be determined much more precisely at points in the wave

oscillation where the fluctuations of the field are smallest. In fact, the fluctuations in the squeezed light field are smaller at some phases than any classical field, so that phase sensors using such a field will be more precise than sensors using classical fields. In fact they will break the standard quantum limit.

This is a costly approach to sensing at present, so it is only used where there is a clear advantage to be had—for instance in the detection of gravity waves by means of very large optical interferometers, such as the GEO 600 project near Hanover in Germany. By using squeezed light, this instrument can detect phase shifts that correspond to a relative path length change of the light equivalent to the size of an atom compared to the distance from the Earth to the Sun.



36. Squeezed light a. has reduced noise in its ffeld amplitude at certain points in its oscillation as compared to laser light b.

Quantum entanglement

Things get even stranger when we consider more than one quantum light beam. Photons can be entwined in such a way that it is impossible to ascribe a property to either of them individually—for example, a colour, position, direction, or

pulse shape. This goes well beyond the fundamental notion of wave- particle duality. It challenges the very notion that in the classical world it is possible to assign real values of properties to physical entities (e.g. in the case of light beams, say frequency and time of arrival, or H- and V-polarization)—in a way that can be revealed by a local measurement in a self-consistent fashion. The fact that this cannot be done for pairs of light beams prepared in certain states, and can be proven so by experiment, is one of the great triumphs of fundamental optical science in the 20th century.

By means of this property, it is possible to use quantum optics to explore the famous conjecture of Einstein, Boris Podolsky, and Nathan Rosen concerning whether a quantum mechanical description of a system of particles can be considered complete, requiring no other information to determine everything about the system. John Bell discovered in the 1960s a means to quantify such a question, and the quest to build an experiment to actually test his hypothesis began in earnest. These are known colloquially as 'Bell tests', and the earliest and currently most convincing work uses pairs of photons, each of which is correlated with the other. It is the nature of these correlations that is so different for quantum particles than for classical ones. It's worth exploring this in a bit more depth in order to get a fuller sense of the strangeness of this quantum effect.

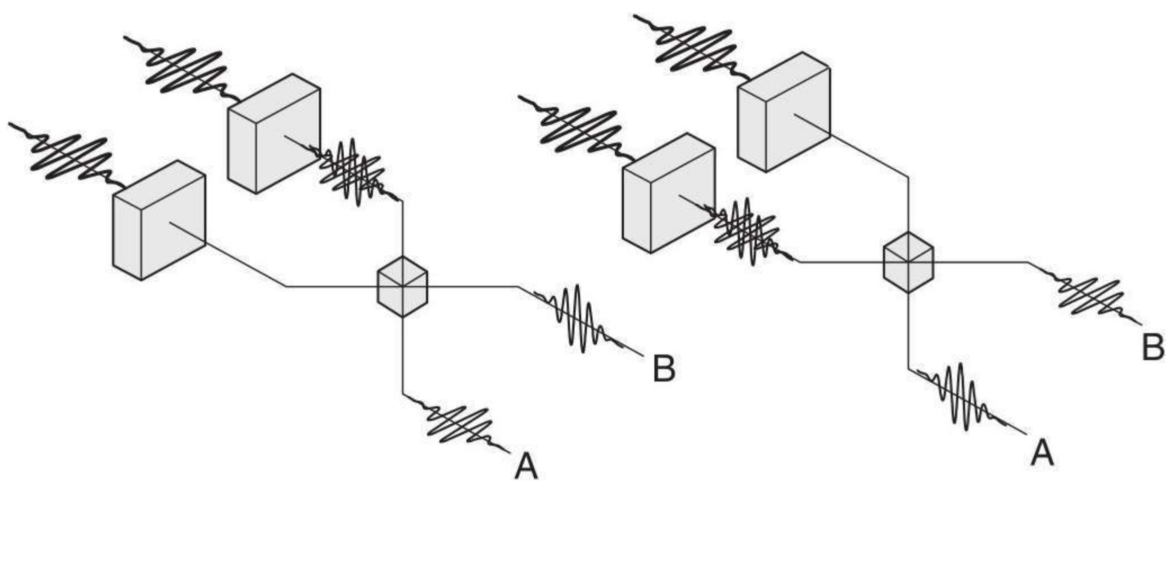
Correlations can be found in almost every situation. Consider for instance the following simple game. A dealer takes two packs of cards, one with green backs and the other with blue backs. The dealer picks one card from each pack and gives one to you and another to your partner. Each of you looks at your card. They always have different colours on the back, of course, but they may have the same colour (red or black) on the front. In fact, you'd expect this to occur half the time, since each of you would expect individually to get either red or black with equal probability (half of the cards in each deck are black and half red).

If you and your partner found that every time you both got red or both got black, you'd say that the cards were 'correlated'. This is about as strong a correlation as you can imagine. In fact, if you both got the same colour more than half the time, you'd also be able to claim the cards were correlated, though clearly the correlations would be 'weaker' than in the first instance. By measuring the correlations, you could determine whether the dealer was cheating, since you might assume she'd start with two independent, complete decks.

We can make an analogy of this kind of correlation for photons using polarization instead of suit for the cards. That is, a horizontally polarized photon might be termed a 'red' photon, and a vertically polarized photon a 'black' one. Then if a source produces photons two at a time, as described above, you can say that it produces correlated

photon beams if it always produces photons with a prescribed polarization, say one vertical and one horizontal, or both horizontal. This type of correlation is termed ‘classical’, since it has a complete analogy to the situation with classical objects like playing cards.

There is a feature of correlations that has an intrinsic quantum mechanical aspect. Let’s say there are two possible states in which the photon pair can be prepared—the first H-polarized and the second V-polarized or vice versa. In the classical world these two situations for two particles are mutually exclusive: either HV or VH is possible, each with a probability of one-half. But, just as the single photon can be in a superposition H and V, so can the pair: HV and VH, shown in Figure 37. This turns out to be a much stronger correlation than is possible with any classical particles, and is called entanglement. It is the most enigmatic property of quantum physics, and has extraordinary consequences.



37. A light source for generating polarization entangled photons.

These are revealed by means of Bell tests. In such a test, you have to consider not only the possibility of correlations in the H and V polarizations of each particle, but also those in the diagonal (D) and anti-diagonal (A) polarizations, each oriented half way between the horizontal and the vertical. (Diagonally polarized light, for example, is shown in Figure 35. Anti-diagonal polarization is oriented at right angles to the D direction.) The analogy with the cards is that you can look at the front of the cards and observe either red (equivalent to H) or black (equivalent to V) suits. Or you could look at the backs and see green (equivalent to D) or blue (equivalent to A).

A quantum game

Now imagine a card game in which the dealer chooses from either pack and gives one card to each player. That means that each player will have a card that could be either red (R) or black (B) on the front (F) and either green (g) or blue (b) on the back (B). The dealer chooses to hand out cards in such a way that if one player looks at the front of his card, and the other the back of hers (F,B), then they never find the result (R,b). Similarly if the first player looks at the back of his card and the other the front of hers (B,F), then they never find the outcome (b,R). However, when they both look at the front of their cards (F,F) they sometimes see (R,R). From this, you would conclude logically that in such a case, had they looked at the front of their

cards (B, B) they would have seen (g, g). That’s what would happen for obviously classical things like cards.

Measurement choice:	Outcomes: Player 1; Player 2; Probability			
F, *F*	R; R; 1/2	R; B; 1/12	B; R; 1/12	B; B; 3/4
F, *B*	R; b; 0	R; g; 1/6	B; b; 2/3	B; g; 1/6
B, *F*	b; R; 0	b; B; 2/3	g; R; 1/6	g; B; 1/6
B, *B*	b; b; 1/3	b; g; 1/3	g; b; 1/3	g; g; 0

38. Table of the probabilities of possible outcomes for a quantum card game.

But in fact, when you take photons (or other particles) that are quantum correlated and do such an experiment it doesn’t turn out that way. What happens is that when the players make measurements of the polarizations using, for the first player, a horizontally oriented polarizer, and, for the second player, a diagonally polarized photon (or vice versa), they never get the results (V = 0, D = 1) and (D = 1, V = 0). Likewise, when they both measure using diagonally oriented polarizers, they sometimes get the result (D = 1, D = 1). Therefore you would logically conclude that when they measured the photons using a horizontally oriented polarizer, they would sometimes get the result (H = 1, H = 1). But, when they do this experiment, they never get this outcome! The table of possible results of such a quantum card game are

shown in Figure 38. Such experiments can, and have been, done using photon pairs. It's not fiction.

Local properties of things

So what is going on? This is the fundamentally weird thing about quantum physics: the conclusion of the quantum card game is that the photons cannot have predetermined values of their polarization when they are prepared at the source. It is as if the cards could not have been of definite suits from a deck with a well-specified card-back colour. This goes against all intuition about cards: they surely have definite properties of a specific suit on the front of each card and a specific colour on the back. No matter whether we know or even the dealer knows what these values are or not, we don't doubt that the cards actually have these properties when they are given to us. And we certainly don't expect that anything we do to them changes those properties. But quantum mechanics tells us that we cannot assign colours to the cards a priori.

It is the measurements that give definiteness in the outcomes. We cannot claim the measurements simply reveal predetermined properties of the photons, which are unknown to the players. It is actually that you cannot assign definite polarizations to the individual photons when they are produced by the source in such a way as to give the outcomes that are actually seen. If you try to devise a way of dealing cards that gives such a result, you'll find that it is

impossible. The cards would need to have the possibility that they can be simultaneously in superpositions of red and black or green and blue, in particular ways. Just so the photons—it is necessary for them to be in superpositions of H and V in a way that gives very specific types of correlations. It is this type of correlation that is termed 'quantum entanglement'.

Entanglement is a very weird concept. It is not possible to find a way to think about it in terms of common everyday objects—as the playing card example was intended to show. Yet entanglement is also very common. It appears in many things at the quantum scale,

even in everyday conditions: in the correlations between electrons in molecules, giving rise to bonds between the atoms making up the molecule, or even relatively small atoms themselves, as well as exotic materials like superconductors.

Surprisingly, entanglement also turns out to have technological implications. You'd hardly think that such an arcane and abstract idea could possibly have any application, but it does. It enables a raft of information processing approaches that cannot be replicated by sending classical waves back and forth. Indeed, the very idea that all information processing systems are at bottom built of something suggests that the design principles of these machines must reflect the underlying physics of the constituent parts—usually classical physics. This has led to

the understanding that basing computing, communications, and measurement on quantum mechanics provides new opportunities for technologies that surpass those of the current generation in unimaginable ways: communications whose security is guaranteed by the laws of nature; computers that can solve ‘uncomputable’ problems; imaging systems that reveal an object that they are not even looking at.

Light plays an important part in implementing such systems. The infrastructure of optical fibre networks, for instance, can be used to distribute random ‘quantum keys’ (random strings of 0s and 1s) completely securely between two parties, which can then be used to encode messages. Such networks can also be used to connect small-scale quantum processors, eventually becoming a distributed quantum computer. Indeed, it has been shown that in principle it is possible to build a quantum computer completely out of light,

though it is extremely challenging to do so. Combining these technologies holds the promise in the future of a quantum Internet,

a radically different way to communicate and process information from the technology we currently use, and all enabled by light.

Chapter 9 Twilight

The atoms of Democritus And Newton’ s particles of light Are sands upon the Red Sea shore Where Israel’ s tents do shine so bright.

William Blake, ‘Mock on, Mock on,

Voltaire, Rousseau’

It is inevitable that in a book of this brevity—covering a topic with as long a history and as broad a range of applications as light—many things simply can’ t be covered. In particular, the fantastic discoveries arising from looking at other regions of the electromagnetic spectrum than the visible, and the ubiquity of optical devices in everyday life are very incompletely described.

The myriad versions of imaging devices, from multi-mirror telescopes (with diameters of tens of metres, and adaptively controlled reffectors to null the efiects of the twinkling sky) to massive synchrotrons (accelerating electrons to the point that they radiate intense X-rays for looking at tiny material structures, both man-made and natural, to reveal the structure of, for instance, biologically important molecules or engineered metals) are only hinted at. But no matter: I hope that you are convinced of the beauty of light, and that a tour through its mysteries and how they were unravelled is as interesting a story as you could want.

Meanwhile, the science and technology of optics is vibrant, opening up new areas for exploration and new applications, often with unexpected fruitfulness. Light continues to reveal new mysteries and to inspire new devices: it still engages and flres our imagination as it has done for centuries.

Further reading

Non-technical books

O. Darrigol, *A History of Optics from Greek Antiquity to the Nineteenth Century*, 2012 (Oxford: Oxford University Press).

H. E. Edgerton and J. R. Killian, *Moments of Vision: The Stroboscopic Revolution in Photography*, 1979 (Cambridge, MA: The MIT Press).

J. P. Harbison and R. E. Nahory, *Lasers: Harnessing the Atom’s Light*, 1997 (New York: Scientiffc American Library, W. H. Freeman & Co).

J. Hecht, *City of Light: The Story of Fiber Optics*, 2004 (New York: Oxford University Press).

J. Hecht, *Beam: The Race to Make the Laser*, 2005 (Oxford: Oxford University Press).

M. Kemp, *The Science of Art: Optical Themes in Western Art from Brunelleschi to Seurat*, 1992 (New Haven: Yale University Press). A. Zajonc, *Catching the Light*, 1993 (Oxford: Oxford University Press).

Specialist texts

M. Bass et al., *Handbook of Optics*, 2000 (New York: McGraw-Hill).

M. Born and E. Wolf, *Principles of Optics*, 7th ed., 1999 (Cambridge: Cambridge University Press).

R. W. Boyd, *Nonlinear Optics*, 3rd ed., 2004 (New York: Academic Press).

J. W. Goodman, *Statistical Optics*, 2000 (New York: Wiley Interscience).

J. W. Goodman, *Introduction to Fourier Optics*, 2004 (New York: Roberts & Co.).

H. Hariharan, *Basics of Interferometry*, 2nd ed., 2003 (New York: Academic Press).

S. Haroche and J.-M. Raimond, *Exploring the Quantum: Atoms, Cavities and Photons*, 2006 (Oxford: Oxford University Press).

E. Hecht, *Optics*, 4th ed., 2001 (Reading, MA: Addison Wesley).

S. M. Hooker and C. Webb, *Laser Physics*, 2010 (Oxford: Oxford University Press).

J. Mertz, *Introduction to Optical Microscopy*, 2010 (New York: Roberts & Co.).

L. Novotny and B. Hecht, *Principles of Nano-optics*, 2nd ed., 2012 (Cambridge: Cambridge University Press).

V. Vedral, *Introduction to Quantum Information Science*, 2006 (Oxford: Oxford University Press).

W. Welford, *Optics*, 3rd ed., 1988 (Oxford: Oxford University Press).

牛津通识课：数字

[英]彼得·希金斯 著
邱明峰 译

东方出版中心

图书在版编目（CIP）数据

牛津通识课. 数字 / (英) 彼得 · 希金斯著 ; 邱明峰译. -- 上海 : 东方出版中心, 2021.1
ISBN 978-7-5473-1186-8

Ⅰ. ①牛… Ⅱ. ①彼… ②邱… Ⅲ. ①科学知识-普及读物②数字-普及读物 Ⅳ. ①Z228②O1-49

中国版本图书馆CIP数据核字（2021）第012100号

上海市版权局著作权合同登记：图字09-2020-1022号

Number: A Very Short Introduction was originally published in English in 2011. This translation is published by arrangement with Oxford University Press. Dook Media Group Ltd is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

Copyright:© Peter M. Higgins 2011

中文版权：©2021 读客文化股份有限公司

牛津通识课：数字

著　　者　[英] 彼得 · 希金斯
译　　者　邱明峰
责任编辑　江彦懿
封面设计　汪芝灵

出版发行　东方出版中心
地　　址　上海市仙霞路345号
邮政编码　200336
电　　话　021-62417400
印 刷 者　天津联城印刷有限公司

开　　本　787mm x 1092mm　1/32
印　　张　7
字　　数　80千字
版　　次　2021年2月第1版
印　　次　2021年2月第1次印刷
定　　价　32.00元

版权所有　侵权必究
如图书有印装质量问题，请寄回本社出版部调换或电话010-87681002联系。

目录

序
01 如何不去考虑数 How Not to Think About Numbers
02 永无穷尽的素数 The Unending Sequence of Primes
03 完美的和不那么完美的数 Perfect and Not So Perfect Numbers
04 密码学：素数的秘密生活 Cryptography: the Secret Life of Pri mes
05 计数的数 Numbers That Count
06 数之冰山的水下部分 Below the Waterl ine of the Number Iceberg
07 向无穷和更远出发! To Infinity and Beyond!
08 并非我们熟知的数 Numbers but Not As We Know Them
术语对照
Contents
Preface
List of illustrations
Chapter 1 How not to think about numbers
Chapter 2 The unending sequence of primes
Chapter 3 Perfect and not so perfect numbers
Chapter 4 Cryptography: the secret life of primes
Chapter 5 Numbers that count
Chapter 6 Below the waterline of the numbericeberg
Chapter 7 To infinity and beyond!
Chapter 8 Numbers but not as we know them
Further reading

序

这本小书致力于用每个人都熟悉的语言，来解释人们所遇到的各种数的定义和它们的行为。数使我们能够将万事万物进行比较。数无处不在，任何对数缺乏理解的人都会在现代世界不知所措。然而，应该意识到，虽然我们已对数习以为常，但它们并非物理实在，而是我们从自己周围的世界抽象出来的概念。为了对它们如何运行给出一个清晰的图像，不如把它们从其他事物中分离出来单独考虑。

本书并非算术温习教程，也不会谈太多有关数的历史。写它的目的是解释数本身以及它们所展现出的行为。看一眼目录就会发现，这本书的前半部分主要关心普通计数用的数，而在后半部分，我们将走出这一范畴：探索从商业和科学中自然产生的问题。这一过程中，对自由计算的需求最终带着我们逐步走进复数的世界。复数是处理大部分数字相关事务的主要框架。这听起来可能有点吓人，不过请放心，真正困难的工作已经有人替你完成了。

现代数的系统并非像大礼包一样从天而降，而是经过若干世纪的发展而来的。人们曾在很长一段时间内迷惑不解。这种情况源于两个根本原因：第一是缺乏一套高效的表示数的方法以便操作；第二则是哲学上的困扰，即如何解释各种不同数的类型以及它们是否有意义。如今，我们对自己在做什么已经有把握得多了，在跟数打交道的时候无需再忧虑这些，因而我们能够在像本书这样短的篇幅内给数的世界描绘一幅完整的图画。这并不是说所有的谜团都已被解开——事实远非如此，你读下去之后就会发现。

彼得·希金斯

科尔切斯特，英格兰，2011

01 如何不去考虑数 How Not to Think About Numbers

我们已经习惯看见写下来的数，也习惯于从中提取出某种意义。然而，一个数字(比如6)同它所代表的那个数并不是同一个东西。就像在罗马数字中，我们会把六[1]这个数写作VI，但是我们意识到这与用现代记号写下的6代表了同一个数，都代表对应6根算筹(IIIIII)的那一类集合。让我们先花一点点时间考虑一下表示和思考数的不同方法吧。

有时候，我们会在无意识的情况下解决一些关于数的问题。例如，假设你正要组织一次会议，想要保证每个人都拿到一份议程。你可以将每份议程逐个标上与会者的名字首字母。只要完成这一工作之前议程一直没用完，你就知道份数是足够的。这样你就解决了问题，而没有用到算术或者直接数数。这里数依然在发挥作用，它们使我们得以将一个集合同另一个集合进行精确比较，即便这两个集合的组成元素有着截然不同的性质。就像上述例子里，一个集合包含了人，而另一个集合则由纸张组成。数让我们可以比较两个集合的大小。

在上例中，你不需要费神去数有多少人将要出席，因为没有必要知道——你的问题是判断议程的份数是否不少于出席人数，而具体的数目无关紧要。但如果你是要为15个人买午饭，你就需要真正数一数人数了。当然，要计算这顿饭总共的开销时，就一定得有人使用算术，哪怕是用计算器求和来得出精确的数值。

现代数字系统让我们能以一种有效且统一的方式来表示数，这方便我们将一个数和其他数做比较，以及在计数问题中进行所需要的算术操作。日常生活中，我们在所有的算术中使用十进制，换句话说，

我们十个十个地数数。这么做的原因是偶然的：我们恰好有十根手指。需要明确的是，让数的系统如此有效的原因并非我们对底数(base)的选择，而是我们在数的表示中使用了**位值进位法**(positional value)，即一个数字的值取决于它在数字串中出现的位置。比如，1984是4份1加上8份10，再加上9份100，再加上1份1000的缩写。

当我们把数写成特定形式的时候，我们想表达什么？理解这一点很重要。在这一章，我们将要考虑数代表了什么，发现不同的计数方法，认识一类非常重要的数(素数)，并且介绍一些找到它们的简单技巧。

人们是如何学会数数的

这里值得我们花上片刻来弄明白构造一个计数系统所需的两个重要阶段。让我们以十进制为例：我们会给孩子们规定两项基本任务，背诵字母表和学习数数。这两个过程表面上看是相似的，却有着本质的不同。英语基于26个字母，简而言之，每个字母对应一个发音，供我们念单词用。总的来说，英语发展的结果是这门语言可以用26个符号来书写。如果不给字母表一个顺序，我们就没法编纂字典。然而，字母表并没有天生的排序。我们所采用并且都曾在学校吟诵过的a，b， c， d…看起来实在很随意。诚然，更常用的字母一般出现在字母表的前半部分，但这也只是粗略的准则，而非铁律。比如，常用字母s和t就很晚才被点名。相比之下，用以计数的数(counting number)，或者叫**自然数**[2](natural number)：1， 2， 3…一旦出现便已经排好了序。例如，符号3用来表示跟在符号2后面的那个数，因此必须列在2的下一位。在一定程度上，我们可以给每个数赋予一个新的符号。但为了处理永无止境的数，我们迟早得放弃不断引入新名字，而只能

开始把它们按批次分组。按十个来分组代表了发展出一个坚实的数字系统的第一阶段，这个方法在古今中外几乎都是一样的。

但是各个文明在细节上却有很多不同。罗马系统除了按十分组外，同样也喜爱用五分组。他们使用特殊符号V和L，来分别代表五和五十。古希腊系统则直接使用了十进制。他们用某些字母来代表数，有时候加上上划线来告诉读者这个符号应被解读为一个数，而不是平常的词语中的一个字母。比如，π 代表80，而γ 代表3，于是他们可以写下π γ 来代表83。与我们的现代数学符号相比，它看起来好像同样高效，也确实跟我们的相差不多，但它们是不一样的。希腊人仍然没能运用进位系统，因为他们每个符号的值都是固定不变的。具体说来，γ π 还是只能表示同样的数——3+80。而如果我们将83里面数字的顺序颠倒，会得到一个不同的数——38。

在印度-阿拉伯系统(Hindu-Arabic system)中数的第二个阶段得以实现。其主要思想是让一个符号的值依赖于它在字符串中出现的位置。这使我们可以只使用一套固定的符号来表示任何数。我们最终选择了由0， 1， 2，…， 9这十个数字组成的数字系统，这套常用的数字系统被称为**十进制**(base ten)。当然，我们完全可以用一个更大或者更小的基本符号集来建立我们的数字系统。我们甚至可以使用两个数字去建立数字系统，比如0和1。这被称为**二进制**(binary system)，在电脑运算中经常被用到。需要明确的是，具有革命性的不是对底数的选择，而是这样一种思想：使用位置来传达额外的信息，从而确定数的值。

例如，当我们写下一个数，比如1905，每个数字的值取决于其在数字串中的位置。这里有5份1，9份100(即10× 10)，以及1份1000(即10× 10× 10)。符号零被用作一个占位符，这很重要。在1905这个例子中，十位并没有贡献，但我们不能忽略它而只写作195，因为那代表了

一个完全不同的数。事实上，每个数字串都代表了一个不同的数，正因为如此，海量的数才可能用很短的字符串表示出来。比如，用不超过10位的字符串便可以给地球上每个人分配一个独特的号码，这样，这个巨大集合中的每一个体都有了个人代号。

不同的古代社会在书写数时使用了不同的底数，但这远不能弥补一个事实，即几乎所有文明都缺少一套真正的进位制系统，更谈不上使用零来作为占位符。在古代世界的所有民族中，巴比伦文明的数字系统是最为接近位值进位法的系统，考虑到他们的古老程度，这实在让人赞叹。然而，他们没有彻底拥抱那个不那么自然的数——零。比如，我们用零来区分830和83，而古巴比伦人刻意回避了在数字串末尾使用这样一个表示空的符号。

意识到零确实是数，这需要跨越一个概念上的障碍。零确实并非一个正数，但它依然是一个数，如果不将它以一种完全自洽的方式吸纳进来，我们的数字系统就是残缺不全的。约公元6世纪，印度迈出了这关键的最后一步。现代的数字系统被称为印度-阿拉伯系统，正因为它是从印度经由阿拉伯传到了欧洲。

有或没有小数的生活

我们将以十为底的位值进位法思想推广到非整数部分，就可以得到我们当今所熟知的完整十进制数字系统(decimal number system)。比如，当我们写下3.14的时候，小数点后的1表示一份$\frac{1}{10}$，类似地，后面的4代表$\frac{4}{100}$。这种两位小数形式的数我们已经很熟悉了，因为我们处理十进制货币的时候，最小单位——取决于具体情况——不是美元、英镑或欧元，而是便士或美分。后者是其主币的百分之一。十进制小数算术是以十为基底的系统的自然推广。在实际应用中，它代表了执行普通加法运算的最佳方法。虽然有这些优点，十进制方法的诞生却并非一蹴而就。一开始它只在数学精英们的小圈子里得到运用，直到

16世纪后半叶才最终进入商业结算和大众使用。甚至在这之后，按照非十的幂次来组织数的方法依然在用。英国直到1971年才采用了十进制货币 。而部分英语地区依然顽固地使用码、英尺和英寸[3]。若要捍卫英制单位，我们可以辩解说它们是围绕人体的尺寸量身定制的。我们的手大约有6～8英寸长，而我们的身高大约是5～6英尺。我们将周围的物体制作得与我们的体型尺寸相似，自然可以方便地用英尺和英寸来丈量它们。但是，如果让1英尺等于10英寸，整个系统也可以很好地运作，还更方便用我们以十为基数的计算方法来处理问题。

为一个数字系统选底数有点像为一张地图选比例尺，这并非基于对象内在的性质，而是类似于赋予它一个坐标系，作为控制用的工具。我们对底数的选择本质上是任意的，面对自然数集1，2， 3，4，…的时候，排他性地使用底数十让我们戴上了有色眼镜。只有掀起这层面纱，我们才能面对面地看清数到底是什么。当我们说起一个数，例如四十九，所有人的脑海中都浮现出两个字符4和9。某种程度上，这对我们谈论的那个数来说不太公平，因为我们毫不犹豫地将49转换成了(4× 10)+9。我们也可以同样轻松地用另一种方式来诠释这个数：49=(4× 12)+1。实际上，在十二进制中，四十九正是被写作41，这里数字4代表4份十二。当然，49这个数最显著的特征是它是7× 7的积。这一特点在七进制中十分明显，此时四十九这个数会被表示为100，这里的1代表一份7× 7。

我们一样有权利为我们的数字系统选择另一个底数，例如12。玛雅人使用了12，而巴比伦人用了60。某种程度上，60是一个计数基底的好选择，因为60有很多因数，它是可以被1到6之间所有数整除的最小的数。不过，选择一个60这样相对大的数作为底数也有缺点，就是需要引入60个不同的符号来表示从零到五十九的所有数。

如果一个数是另一个数的整数倍，则称第二个数是第一个的**因数**[4](factor)。比如，6是42=6× 7的一个因数；但8不是28的一个因数，这是因为28含有三个8还余4。数字系统的底数拥有众多因数会是一个很方便的性质。这就是为什么相对于我们的数字系统来说，底数选择十二也许比十更好，因为12的因数是1， 2， 3， 4， 6和12，而10只能被1， 2， 5和10整除。

数字系统的有效性，加上我们对它高度熟悉，给了我们一种虚假的信心，同时也带来了一定的局限。我们会更愿意使用单个的数，而非一个算术表达式。例如，大多数人情愿谈论5969，而不是47× 127，即使这两种表示方式代表了同一个东西。仅在“求出最终答案”5969之后，我们才觉得我们“有”了这个数，从而可以正视它。然而，这里面却包含着一丝虚妄的成分，因为我们只是将这个数写成了多个十的几次幂的和。若是将这个数分解为一系列因数的乘积，从这个等价形式中，我们反而可以更好地推断它一般意义上的构造和其他的性质。确切地说，5969这个标准形式能让我们将它和其他以同样方法表示的数直接比较大小，但并不能展现出这个数的全部特性。分解为因数的形式可能有用得多。在第4章中，你会发现其中的一个原因，即一个数的十进制表示可能掩盖了关键的因数。

古人比我们多拥有的一项优势是，他们并未被十进制的思路所束缚。谈到数的规律时，他们会自然地想到一个数可能拥有或缺乏某些特殊的几何性质。比如，10和15这样的数是**三角的**(triangular)。这可以通过图像展现在我们面前，正如十瓶制保龄球中排成三角形的球瓶，以及斯诺克台球中15个红球摆成的三角。但如果只用十进制展示这些数，这样的画面就不会出现在我们眼前。我们要放下十进制的思维定势，告诉自己可以从很多不同角度来思考数，才能重新获得古代人天然就拥有的自由。

这样，把我们自己解放出来之后，我们可能会选择重点关注一个数的**因数分解**(factorization)——把数写成一些更小的数的积。因数分解揭示了一个数内部构造的相关信息。通常，我们仅将数看作科学和商业的仆人，如果能把这个习惯暂时丢在一边，花一点点时间专门研究数本身，而不与任何其他事物做关联，我们就能发现很多原来隐藏着的信息。单个数的性质可能会在自然界中以有序的模式展现出来，这比单纯的三角形或方形来得更微妙，比如，向日葵螺旋形的头状花序代表的斐波那契数。我们将在第5章中介绍这种类型的数。

素数数列概览

数的荣耀中有一项是如此显而易见，以至于很容易被忽视——它们每一个都是独一无二的。每个数有自己的结构——如果你喜欢的话，可以称之为个性。单个数的个性非常重要，这是因为当一个特定的数出现时，它的特征会反映在它所描述的集合的结构中。当我们执行加法和乘法这些基本的数的运算时，数之间的关系也会显现出来。显然，任何比1大的自然数都可以表示成更小的数的和。但是，当我们开始将数相乘，我们很快注意到有些数从来都不会在我们得到的乘积里出现。这些数就是**素数**[5](prime number)。它们代表了乘法的构成要素。

一个素数是一个像7，23或103这样的数。它有且仅有两个因数——1和它本身。我们并不把1看作一个素数，因为它只有1个因数。那么第一个素数便是2，这也是仅有的偶素数，接下来的3个素数3，5和7都是奇数。大于1而又不是素数的数称为**合数**(composite)，因为它们由一些更小的数复合而成。数4=2× 2=2[6]是第一个合数；9是第一个奇合数，9=32也是一个**平方数**(square)；6=2× 3是我们第一个真正意义上复合的数，因为它由两个不同的大于1的因数复合而成。而$8=2^3$是第一个**立方数**(cube)，立方数是指等于某个数的3次幂的数2。

紧接着个位数的是我们选择的底数10=2× 5，它本身就是特殊的，而且还是三角的，因为10=1+2+3+4(回想一下十瓶制保龄球)。接下来我们有一对**孪生素数**(twin primes)11和13，它们是由12隔开的两个相邻的奇数，且同为素数。考虑到数值的大小，12显得有很多因数。的确，12是第一个所谓的**盈数**[7](abundant number)，因为它的**真因数**(proper factors)——严格小于它自身的因数——之和超过了这个数本身：1+2+3+4+6=16。数14=2× 7可能看起来并没有什么特别的，但这实在是一条悖论，作为**第一个**不特别的数本身就是一件特别的事。到了15=3× 5，我们又遇到了一个三角数，并且它是第一个等于两个真因数之积的奇数。数$16=2^4$不仅是一个平方数，还是第一个四次幂数(在1以后)，这使得它十分特别。17和19是另一对孪生素数。对于18和20，以及更多数的独有性质，我就留给读者朋友自己去观察了。对每一个数，你都可以说它是特别的。

回到素数，它们中的前20个是：

2，3，5，7，11，13，17，19，23，29，31，37，41，43，47，53，59，61，67，71。

显然，在数列的开始部分，素数是十分常见的，因为小的数没有多大可能会有因数分解。在这之后，素数越来越稀少。比如，只有一组连续3个奇素数：3，5，7。这三者的组合是独一无二的，因为每3个奇数就会出现1个3的倍数，所以这种情况再也不会发生。此外，素数出现频率减少的过程是十分松散的，还惊人地不规律。比如，30～40之间只有两个素数，即31和37，但刚过100就会有两个“相继的”孪生素数对，即101，103和107，109。

素数在数千年来一直深深吸引着人们[8]，因为它们无穷无尽(我们在下一章会证明这个说法)，在自然数中现身的方式又神出鬼没。它

们性质中神秘莫测的这一面在现代密码学(cryptography)中被加以运用，来保护互联网上的机密信息。这将是第4章的话题。

素性检验：素数整除性测试

为了找出不大于某个数(比如100)的所有素数，最容易想到的方法是把所有数写下来，再寻找并划掉里面的合数。基于这一思想的标准方法叫作埃拉托斯特尼筛法[9](Sieve of Eratosthenes)。方法如下：首先圈出2并划掉列表里2的倍数(即其他偶数)。然后回到开始，圈出第一个尚未被划掉的数(即3)，划掉剩下数表里它的所有倍数。重复这一过程足够多遍，剩下没有被划掉的就是素数。即便它们中一些被圈出来了，而另一些没有。例如，图1展示了对不大于60的数运用筛法的过程。

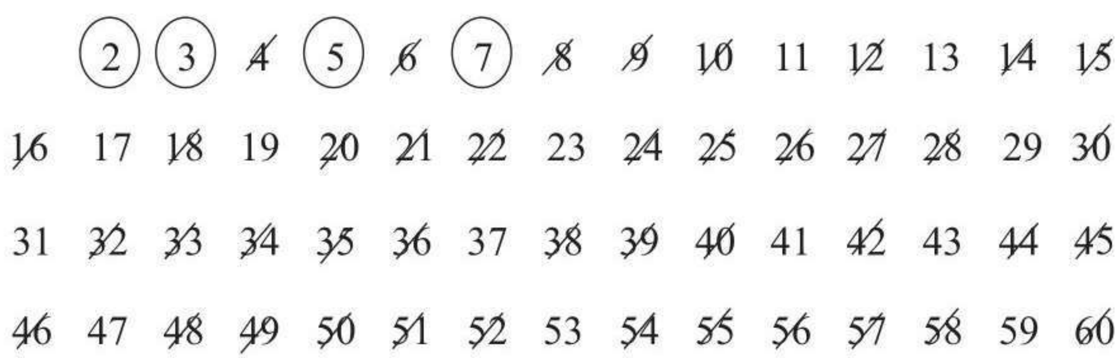


图1　素数筛：没有被划掉的数即为60以内的素数

你怎么知道什么时候该停止筛选呢？重复这一过程，直至圈到一个数，大于整张表中最大数的平方根。比如，当你在不大于120的数中筛选素数时，需要筛掉2， 3， 5和7的倍数，接着当你圈出11，就可以停下了。因为$11^2=121$。此时，你已经圈到了第一个大于你的最大数(这里是120)的平方根的数，剩下的素数虽未被触及，但所有的合数都已经被划掉，它们都是2， 3， 5或7的倍数。

不难看出为什么列表中最大数n的平方根决定了需要筛选的遍数。任何列出的合数m都至少有一个素因数，它**最小的**素因数必小于n的平方根，因为两个或两个以上大于$\sqrt{n}$的数的积大于n(并且因此也大于m)[10]。

素性问题的另一面是：任一给定数n是素数还是合数。为了确定这一点，我们可以用小于$\sqrt{n}$的素数逐个除n，如果n不能被所有这些数整除，则其为素数，否则不为素数。鉴于这个原因，对于小素数，如2，3，5，7，…了解一些简单的整除性测试方法会带来很多便捷。下文的方法让这个需求很容易得到满足。

检测能否被2或5整除很容易，因为这两个素数是我们的底数10的素因数。看出这一点，你只需查看待检数n的最后一位：当且仅当个位为偶(即0，2，4，6或8)时n可以被2整除，当且仅当n最后一位为0或5时它含有因数5。不管数n有多少位，判断n是否为2或5的倍数，我们都只需检查最后一位。对于不能整除10的素因数，我们需要多做一点工作，尽管如此，仍然有一些整除性测试方法，比计算完整的除法更快捷。

一个数能被3整除，当且仅当其各位数字之和能被3整除。例如，数n= 145 373 270 099 876 790的各位数字之和是87，而87=3× 29，因此n可被3整除。当然，我们还可以将这个测试应用于数87，然后继续在每一阶段都求出各位数字之和，直到结果显而易见。对于给出的例子这样做，可以得到以下数列：

$$145\ 373\ 270\ 099\ 876\ 790 \to 87 \to 15 \to 6 = 2 \times 3.$$

你会发现，这里列出的所有整除性测试方法都如此快捷，你可以相对轻松地处理有几十位的数，甚至比手持式计算器能处理的最大的数还要大上数十亿倍。

下面要给出不大于20的剩下所有素数的整除性测试方法，因为它们都属于同一个类型。虽然它们成立的原因不那么明显，但是这些方法应用起来都很简单。即便这里没有收录论证，想证明它们的正确性并不是特别困难。

让我们从给定一个数n是否能被7整除开始。将n的最后一位数字乘以2得到一个数，n移除最后一位数字之后的结果再减去这个数。当n是7的倍数时，新的数也恰好会是7的倍数。我们重复这一过程，直到结果变得显然。例如，设$n=3465$，5乘以2得10，因此我们从346减去10，得336；接着再做一次，将6乘以2得12，从33减去12得$21=3\times 7$，因此n可以被7整除。如果你忘记了7的乘法表，我们可以再次进行，从2中减去2倍的1得0，而0可以被7整除，因为$\frac{0}{7}=0$。(可以用0除以一个非0的数，但反过来的话——除以0——则没有意义。)甚至大如好几千万的数也可以用这个方法简单地处理。在本例和之后的例子中，我们仅仅列出**算法**(algorithm)在每个阶段输出的数。算法是指用来解决一类特定问题的一个机械的过程。正如我们这里用到的方法：

$$n=27\ 916\ 924\rightarrow 2\ 791\ 684\rightarrow 279\ 160$$

$$\rightarrow 27\ 916\rightarrow 2779\rightarrow 259\rightarrow 7\ .$$

因此n可以被7整除。每执行一次指令循环，我们都至少能减少一位数，因此执行循环的次数大约与初始数的长度相同。

为了检测n是否有因数11，将原数的最后一位移除，再从新数中减去原数的最后一位，依此重复。比如，我们的方法揭示了下面这个数是11的倍数：

$$4\ 959\ 746 \to 495\ 968 \to 49\ 588$$

$$\to 4950 \to 495 \to 44 = 4 \times 11 .$$

检测能否被13整除，将原数最后一位移除，再用新数**加上**原数最后一位的4倍，接着类似于7和11的方法，不断重复。比如，13是下面这个数的一个素因数：

$$11\ 264\ 331 \to 1\ 126\ 437 \to 112\ 671$$

$$\to 112\ 71 \to 1131 \to 117 \to 39 = 3 \times 13 .$$

接下来是17和19。对于17我们将原数最后一位移除，再用新数减去原数最后一位的5倍，重复操作直到可判断整除性；对于19我们将原数最后一位移除，再用新数加上原数最后一位的两倍，重复操作直到可判断整除性。例如，我们来检测18 905是否能被17整除：

$$18\ 905 \to 1865 \to 161 \to 11 .$$

因而它不是17的倍数。但对于19，整除性测试给出了另一种结论：

$$18\ 905 \to 1900 = 100 \times 19 .$$

拥有了这一组测试，你现在就可以检查不大于500的所有数的素性，因为23^2=529超过了500，因此19是你需要考虑的最大的素因数。

例如，为了确定247的性质，我们只需要检查它是否能被不大于13的素因数整除，因为下一个素数的平方，$17^2=289$，超过了247。运用对13的测试，我们发现247→(24+28)=52→13，这是一个13的倍数(247 = 19× 13)。

素数相乘可以得到无平方因数(square-free)的数，要想构造关于这些数的整除性测试，可以叠加并行其素数因子的整除性测试。比如，对于42=2× 3× 7，一个数n能被42整除，当且仅当n可以通过2，3和7的三项整除性测试。但是，对于那些有平方数因数的数，像$9=3^2$，则不能由此得到。顺便说一句，9是n的因数当且仅当9也是n的各位数字之和的因数。

你也许会问：数千年以来，那些聪明的数学家们难道还没有想出更好、更精妙的检测素性的方法？答案是有的。2002年，我们发现了一个相对快速的判断一个给定的数是否为素数的方法。不过，如果给定的数恰好是一个合数，那么这个所谓的“AKS素性测试”并不能给出该数的因数分解。虽然原则上说，找到一个给定数的素因数的问题可以通过试错来解决，但实际操作中，这对于很大的整数依然难以实现。正因为此，它构成了很多互联网加密方法的基础。我们会在第4章回到这个话题。在那之前，我们会在接下来的两章中更近距离地认识一下素数和因数分解。

[1] 这里的中文数字“六”强调的是“六”这个数本身的概念。

[2] 我国的自然数包含零，而本书的自然数不包含零。

[3] 1码等于3英尺等于36英寸。

[4] 或称因子、约数，英文也可以写作divisor。

[5] 英文也可以写作prime。

[6] 这里指的立方数从2的立方开始，因为$1=1^3$。

[7] 又称丰数或过剩数。

[8] 最近的例子是孪生素数猜想，传奇美籍华裔数学家张益唐在2013年取得重大进展，引起轰动。

[9] 又称埃氏筛或素数筛，简称筛法。埃拉托斯特尼(公元前276—公元前194)，古希腊数学家、地理学家。

[10] 当解释有关任意数的性质的时候，数学家们会用符号为讨论的对象赋予名称。对于数，这些名字通常都是小写字母，如m和n；两个数的积m× n经常简写为mn。

02 永无穷尽的素数 The Unending Sequence of Primes

镶嵌在数的拼图中的素数

我们怎样才能确定素数不会越来越稀少，最终逐渐消失殆尽呢？你可能会认为由于有无穷多自然数，而每一个都可以被分解为素数的乘积(这一点我们一会儿仔细解释)，那么必然得有无穷多个素数才能承担这一工作。虽然这个结论是正确的，但它并不能从上述观察中得出。这是因为如果我们从有限个素数开始，仅使用这些给定的素因数，我们就能制造出无穷多不同的数。确实如此，任何单个素数都有无穷多个幂次。比如，素数2的幂分别为2， 4， 8， 16， 32， 64，…因而完全可以设想：只有有限多的素数，**每个**数都是那些素数的幂的乘积。更糟的是，我们能构造出一个给定数的任意长度的幂数列，或它的任意多倍数列，却没法用同样的手段构造出一个由不同素数组成的无穷数列。对于素数，我们还是得去搜寻，到底怎么才能确定它们不会绝迹？

在这一章结束的时候，我们便都会对这一点确定不疑了。首先，请你注意素数的一个值得一提的简单“规律”。除了2和3以外的每一个素数，都与一个6的倍数相邻。换句话说，在这两个数之后的每一个素数，都是像$6n\pm 1$这样的形式，这里n是某个确定的数。(记住6n是$6\times n$的缩写，符号± 意思是加或减。)原因很好理解：每个数都一定可以写成以下6种形式中的一种：6n，$6n\pm 1$， $6n\pm 2$， $6n+3$，因为没有数与6的某个倍数距离超过3。例如，$17=(6\times 3)-1$， $28=(6\times 5)-2$，$57=(6\times 9)+3$。事实上，这6个形式的数是循环出现的，这意味着当你写下**任意**6个连续的数，6种形式每个都会出现且仅出现1次。在这之后它们会一遍又一遍地循环出现，并且出现的顺序总是相同的。很显

然6n和6n+2形式的数都是偶数。而任何形如6n+3的数都可以被3整除。因此，除了很显然的例外2和3，只有形如6n± 1的数可能是素数。如果6n± 1**两者都**是素数，这种情况恰好对应于孪生素数：比如(6× 18)± 1给出一对数107和109，我们在第1章中提到过它们。你可能会猜测每对6n± 1中**至少有1个**是素数——这对于100以下的素数来说的确是对的，但往前不远就存在着第一个例外：(6× 20)-1=119=7× 17，(6× 20)+1=121=11× 11。当n = 20时这两个数都不是素数。

素数之所以重要，主要是因为每个数都可以写成一系列素数的乘积，并且这么做的方法本质上只有一种。为了找到这个特别的分解，我们只需用某种方法分解这个给定的数，接着继续分解在因数中出现的合数，直到这样的分解不能再继续为止。比如，我们可以说120=2× 60，接着继续将合数因数60分解：

120=2× 60=2× (2× 30)=2× 2× (2× 15)=2× 2× 2× 3× 5.

我们说120的**素因数分解**(prime factorization)为23× 3× 5我们当然也可以用另一种途径得到这个结论。比如：

120=12× 10=(3× 4)× (2× 5)=(3× (2× 2))× (2× 5).

但如果将素因数从小到大重新排列，我们依然得到了与之前相同的结果：120=2^3× 3× 5。

至少在上面这个例子中分解的唯一性是真的。你可能或多或少也已经熟悉数的这一性质，但是如何保证这个结论适用于每一个数？任何数都可以分解为素数的乘积，这已经足够清楚了。但是，一般来说有不止一种办法可以完成这个任务，那么我们如何确定这个过程总能给出相同的最终结果呢？这是一个重要的问题，因此我将花上一点时间概述一下推理的过程，从而使我们能绝对信赖这个结论的正确性。

这个结果来自于素数的另一个特殊性质，我们叫它**欧几里得性质**(Euclidean Property)：假设有一个由两个或更多数相乘得到的积，如果一个素数是该乘积的一个因数，那么它也是构成这个乘积的某个因数的因数。比如，7是8× 35=280(也可以看作乘积280=7× 40)的一个因数，同时我们注意到7也是35的一个因数。这个性质刻画了素数的特征，因为没有合数能够保证同样的结论成立。例如，我们可以看出6是8× 15=120(也可以看作120=6× 20)的一个因数，6却不是8或15的因数。

素数总是拥有以上性质，这一事实可以用被称为欧几里得算法[1](Euclidean Algorithm)的推理来证明。我们将在第4章解释这个算法。如果我们暂且相信这个结果，那么就不难解释为什么不存在一个数拥有两个不同的素因数分解。假设存在拥有两个不同素因数分解的数，那么就存在一个最小的这样的数，让我们用n表示它。n有两种素因数分解。当把n的素因数从小到大排列的时候，这两个分解不相同。我们要证明这是矛盾的，因而假设一定为假。

倘若n的这两种分解有一个共同的素数p，我们可以从两个中都消去p，从而得到一个更小的数$\frac{n}{p}$的两种不同的素因数分解。由于n已经是最小的有两种相异素数分解的数，这是不可能的。因此，n的不同分解中含有的素数集合必定没有交集。现在，取n的一种分解中的一个素数p。因为这个素数p是n的一个因数，可以推断p也是第二种分解的一个因数。所以，根据欧几里得性质，p是第二种分解中某个素数q的一个因数。**然而，由于p和q均为素数**，这仅在p=q时是可能的。但是我们已然排除了这种情况，因为n的这两种分解没有共同素因数。于是我们发现了矛盾，说明这样的n是不可能存在的。综上得证，任何数的素因数分解都是唯一的。

值得一提的是，如果我们把数1包括在素数里，素因数分解的唯一性将不再成立。这是因为我们可以将1的任意幂次方乘上一个分解式，最后的积保持不变。这表明1和素数们有着本质上的不同。基于此，把1排除在素数的定义之外是正确的。

欧几里得：素数的无穷性

让我们回到这个问题：我们怎么知道素数无穷无尽，没有办法找到最大的素数？如果有人声称101是最大的素数，你即刻便能反驳他，因为你可以证明103没有因数(除了1和103以外)，因此103是一个更大的素数。接下来你的朋友可能会承认自己大意了，他应该说103是所有素数中最大的。这时候你还可以指出107也是一个素数，从而再次表明他错了。然而你的朋友可能还是执迷不悟，他搬出你们所知的最大的素数。他甚至可能退一小步，承认自己并不知道最大的素数是多少，却继续说他很确定有这么一个数。

解决这个问题最好的方法是：对于任何假想的有限的素数集合，证明我们都能给出一个不在这个集合中的新的素数。比方说，如果有人号称在某个位置存在一个最大的奇数，你可以反驳说，假如n是奇数，那么$n+2$是一个更大的奇数，因而不可能有一个最大的奇数。然而，这个方法对于素数来说可就不那么简单了——给定一串有限的素数，我们没有办法使用这个集合来造出一个素数，并且表明它比集合中所有的数都大。那么，或许就真的有一个最大的素数？我们怎么才能知道那位固执的朋友是不对的呢？

欧几里得是知道的。约公元前300年，希腊有个数学家叫亚历山大里亚的欧几里得(Euclid of Alexandria)。他是一切作为定语使用的词“欧几里得”(Euclidean)本人。从给定一个集合p_1，p_2，…，p_k(其中每个p_i表示一个不同的素数)，他也没能找到生成一个新的素数的途径，于是他退回一步，找到了一个更加微妙的论证方法。他证明了在

某个给定的数的范围内，总是存在一个或更多的新素数(但他的论据并不足以让我们在那个范围内精确地定位素数)。

他的证明如下：设p_1，p_2，…，p_k为前k个素数。考虑比所有这些素数的积还大1的数n，即$n=p_1p_2\cdots p_k+1$。n要么是一个素数，要么可以被一个小于它自身的素数整除。如果是后一种情况，这个素因数不可能是p_1，p_2，…，p_k中的任意一个。因为假设p是p_1，p_2，…，p_k中任意一个数，将n除以p会有余数1。于是可推知n的任何素因数都会是一个新的素数，并且比p_1，p_2，…，p_k里面所有的素数都大，但不超过n自己。特别是，由此可知不存在任何包含所有素数的有限的列表，因而素数数列会不断延续下去，永不终结。欧几里得关于素数无穷性的证明是永恒不朽的，是数学中最受敬仰的证明之一。

虽然欧几里得的推理没有确切地说明在哪里能找到下一个素数，但现在我们对素数出现的频率已经取得了深入的理解。例如，如果我们任意取两个没有公因数的数，比如a和b，并考虑数列a，$a+b$，$a+2b$，$a+3b$，…，德国数学家约翰·狄利克雷(Johann Dirichlet)证明这样一个数列包含无穷多个素数。(当然，如果a和b有公因数——比方说d，这就没有希望成立了。因为如此一来，列表里每一个成员都是d的倍数，因而不是素数。)当$a=1$，$b=2$时，我们得到奇数组成的数列。由欧几里得的证明，我们知道它包含无穷多素数。实际上，通过对欧几里得的方法进行一些十分简单的改进，还可以证明其他一些特殊情况，比方说以下形式的数列：$3+4n$，$5+6n$，$5+8n$(n依次取1，2，3，…这些值)，它们各自都含有无穷多素数。但是，要证明狄利克雷的一般结论就非常难了。

另一个可以简单陈述的结果是，对于任意给定的数n($n\geqslant 2$)，至少存在一个素数大于n但小于$2n$。这在历史上被称为**伯特兰猜想**[2](Bertrand’s Postulate)，它可以用非常基本的数学知识来证明，虽

然这个证明本身有些取巧。我们可以使用下面列出的素数，对取值不大于4000的n验证这个论断。首先观察到这个列表中，位于打头的素数2之后的每个数都小于前一个数的2倍：

2，3，5，7，13，23，43，83，163，317，631，1259，2503，4001.

对于每个不大于4000的n，取上表中不超过n的最大素数p，它的后面一个素数q即位于$n<q<2n$的范围中。这就保证了伯特兰-切比雪夫定理对于不大于4000的所有n都是成立的。例如，当$n=100$，$p=83$，那么$q=163<2\times 100$。再使用一条巧妙的论证，这个论证涉及所谓的中央二项式系数(central binomial coefficient，将在第5章中介绍)，还可以证明这个定理对大于4000的n也是正确的。

然而，我们不用走太远，就又会发现似曾相识却尚未解决的问题。举例来说，没有人知道是不是在两个连续的平方数中间总存在着一个素数。另一个观察是似乎存在足够的素数，从而可以保证每个大于2的偶数n总可以写成两个素数之和(**哥德巴赫猜想**，Goldbach’s Conjecture)。对于n小于10^{18}的情况，这个结论已经被直接验证。自然，我们可能会寄希望于沿着伯特兰-切比雪夫定理的思路找到一个证明。在大于某个给定的整数N时，基于对素数分布的已有知识，我们试图证明：对于任何偶数$2n\geq N$，至少存在一对素数p，q，构成方程$p+q=2n$的解。这个途径迄今还没能成功，不过这些思路产生了一些弱一点的结果。比如，1939年之后我们知道了：每个足够大的奇数是至多三个素数的和；每个偶数是不超过300 000个素数的和。但要想完整地证明哥德巴赫猜想，似乎还有很长的路要走。

还有一个简单的结果，颇有一些上面介绍的这类论断的味道。它说的是：存在一个小于40亿的数n，有十种不同的方法，可以将它写成四个不同的立方数之和。已知$1729=1^3+12^3=9^3+10^3$是最小的能用两种

方法写成**两个**立方数之和的数。不过，要想知道一个数n存在，我们并不一定非要确定它的大小。有时候可以明确地知道一个问题有解，而不是精准地找到一个解。

在这个例子中，我们先指出如果取四个不同的数，它们都不大于一个固定的数m，那么求它们的立方和，结果将小于$4m^3$。同时，倘若m=1000，那么通过简单的计算就可以发现，选四个不同的数求其立方和，所有可能的情况已经超过了$4m^3\times 10$种。由此可推出存在**某个**数$n\leqslant 4m^3=4\ 000\ 000\ 000$一定可以写成四个立方数的和，且至少有十种不同的写法。具体的计算涉及二项式系数(将在第5章中介绍)，但并不困难。

卡尔·弗里德里希·高斯(Karl Friedrich Gauss)是19世纪引领世界的德国数学家和物理学家。素数分布的全景图像可以用他的观察来概括：对不大于数n的素数的个数p(n)，可以由$\frac{n}{\ln n}$近似地给出，并且这个近似随着n的增大而越来越精确。例如，如果我们取n为100万，比值$\frac{n}{\ln n}$告诉我们在这一阶段，大约每12.7个数中就会有一个素数。高斯对其中某些细节还给出了更精确的观察结果，他的猜想直到1896年才被证明。这里提到的对数函数是所谓的**自然对数**(natural logarithm)，它的值不是以10为底数的幂的指数，而是以一个特殊的数e为底数的幂的指数。e大约等于2.718。关于这个著名的数e，我们在第6章中会听到更多的故事。

数论中最著名的悬而未决的问题是黎曼猜想[3](Riemann Hypothesis)，要阐述它必须用到复数(complex number)——我们还没有介绍到。在这里提到它，是因为可以通过素因数分解的唯一性，重新表述这个问题的对象，使得新的提法中出现了一个包含所有素数的无穷乘积。借助这个表述我们发现，这个猜想表明，素数整体上的分布符合一条规律，那就是在大范围内，素性的出现是随机的。当然，

某个数是否为素数不是一个随机事件。猜想里说的是，就很大的范围而言，素性是随机显现的，没有任何其他的规律或者结构可循。很多数论学家衷心希望，在其有生之年，这条有150年历史的猜想能有个定论。

素数是一个极其自然的数列，以至于我们会几乎无法抗拒地去搜寻它们的规律。然而，却不存在有关素数的真正有用的公式。也就是说，没有已知的规则能够生成所有的素数，甚至无法计算出一个完全由不同的素数组成的数列。存在一些形式简洁的公式，但几乎没有实用价值，要计算其中一些的值甚至需要素数相关的知识，因此本质上它们算是作弊。形如n^2+n+41的表达式称作**多项式**(polynomials)，这一个多项式能够产生极其大量的素数。例如，让n依次取值1，7和20，会分别得出素数43， 107和461。确实，从n=0到n=39的所有输入值，这一表达式的输出都是素数。但当取n=41时，这个多项式就令我们大失所望了，因为结果会有因数41。并且，对于n=40也失败了，因为

$$40^2+40+41=40(40+1)+41$$

$$=40\times 41+41=(40+1)41=41^2.$$

可以轻而易举地证明，一般而言没有某个多项式能给出一个素数的公式，即便允许表达式中存在高于2的幂次也不行。

的确有可能设计出仅用一两句话描述的素性检验的方法。但是，想要它们有用，还需要再快捷一点，至少在某些情况下得快于第1章中描述的直接验证方法。一个有名的结论叫作威尔逊定理(Wilson’s Theorem)。它的表述使用到了一类叫作**阶乘**(factorial)的数，我们将在第5章里再次与它相见。数n!，读作“n的阶乘”，就是所有不大于n的正整数的乘积。比如，$5!=5\times 4\times 3\times 2\times 1=120$。威尔逊定理便是

一条极其简洁的陈述：当且仅当p是$1+(p-1)!$的一个因数时，数p为素数。

这个结果的证明不是很困难，实际上，其中的一个证明方向是很明显的：如果p是合数，比如$p=ab$，那么由于a和b都比p小，它们都会是$(p-1)!$的因数。因而p也是这个阶乘的一个因数。由此推出当我们用$1+(p-1)!$除以p时，会得到余数1。(a=b的情况需要多考虑一点点。)这很容易让人想起欧几里得对素数无穷性的证明。于是可知**如果**p是$1+(p-1)!$的一个因数，**那么**p必为素数。反过来的结论证明起来会难一点：**如果**p是素数，**那么**p是$1+(p-1)!$的一个因数。但这才是定理真正令人惊讶的方向，不过读者朋友可以轻松地验证一些特殊情况，比如，素数5确实是1+4! =1+24=25的一个因数。

威尔逊定理所做的，是将判定p是否为素数的问题，从一系列除法问题(检验被所有不超过$\sqrt{p}$的素数除)转化为一个单独的除法问题。然而，被除数$1+(p-1)!$(即使对于很小的p)也是巨大的。虽然威尔逊定理是一条简明的规则，但它在识别特定素数上并没有进入实际应用。举个例子，用威尔逊法检验13是素数，要求我们验证13是1+12! =479 001 601的因数。(如果使用第1章中13的整除性测试，读者朋友可以验证威尔逊是对的!)拿这个对比一下仅检验13既不能被2也不能被3整除的工作量吧。虽然威尔逊定理在素数检验中没什么价值，但比装饰价值还是要强一点，它可以用来证明其他理论结果。

最后，基于定义，阶乘含有很多因数。我们可以利用这一点证明，不存在**仅含**素数的**算术数列**[4](arithmetic sequence)，或者说仅含素数且形如a，$a+b$，$a+2b$，$a+3b$，…的数列。因为可以证明相邻素数的间距可以任意地大，而前述数列相邻元素之差始终固定为b。想要理解这一点，考虑由n个连续整数构成的数列：

$(n+1)!+2$，$(n+1)!+3$，$(n+1)!+4$，…，$(n+1)!+n+1$.

这些数每个都是合数，因为第一个可以被2整除(两项都含有因数2)，第二个可以被3整除，下一个可以被4整除，以此类推，直到最后一个——含有因数$n+1$。因而对于任意给定的n，我们都有一串 由n个连续整数组成的数列，其中的每一个数都不是素数。

在下一章中，我们将不再聚焦于具有最少因数的数(素数)，而是转向拥有很多因数的数。不过我们会发现，它们和一些非常特殊的素数之间存在着令人惊讶的联系。

[1] 在中国一般称之为辗转相除法。

[2] 该猜想由伯特兰提出，后由切比雷夫证明。约瑟夫·伯特兰(Joseph Bertrand)，法国数学家。巴夫尼提·列波维奇·切比雪夫(Pafnuty Lvovich Chebyshev)，俄罗斯数学家。

[3] 格奥尔格·弗雷德里希·伯恩哈德·黎曼(Georg Friedrich Bernhard Riemann)，德国数学家。

[4] 即等差数列。

03 完美的和不那么完美的数 Perfect and Not So Perfect Numbers

数的完美性

对于取值小的数，我们通常能轻易找到特殊的性质来刻画它们，比如，3是唯一等于之前所有数之和的数，而2是仅有的偶素数(这使得它成为最怪异的素数)。6这个数有个独一无二的性质，它既是所有小于自身的因数的和，也是它们的乘积：6=1+2+3=1×2×3。

毕达哥拉斯学派(Pythagoreans)将6这样的数称作**完美的**[1](perfect)，意思是这个数是其所有真因数之和。对于一个数，我们把严格小于这个数本身的因数叫作它的真因数。这种完美性着实非常罕见。前5个完美数是6， 28， 496， 8128和33 550 336。对于这些偶的完美数我们已经了解了很多，然而直至今日，依然没有人能回答古代人提出的基本问题，即是否有无穷多个这类特殊的数。另外，没有人找到过一个奇的完美数，也没有证明其不存在。任何奇完美数必然极其地大，并且由于奇完美性，这个数必须满足一长串 特殊的性质。但是，所有这些限制条件还不足以排除这样一个数存在的可能——可以想象，这些特殊性质会引导我们去搜寻还未曾现身的第一个奇完美数，它可能只是在等着被发现。

欧几里得早就发现，偶完美数与一列非常特殊的素数有紧密的联系。它们被称为**梅森素数**(Mersenne primes)，是以17世纪的法国教士马兰·梅森(Marin Mersenne)命名的。

梅森数(Mersenne number)是形如2^p-1的数，这里的p是一个素数。举个例子，如果你取前四个素数2，3，5和7，那么可以看出前四个梅森数是：3，7，31和127。读者朋友可以很快验证它们都是素数。

如果p非素，比方说$p=ab$，那么$m=2^p-1$当然也不是素数，因为可以验证在这种情况下m含有因数2^a-1。倘若p为素，则对应的梅森数常常是素数，至少在我们看来是这样的。

早在公元前300年，欧几里得就阐释过：一旦你有一个素的梅森数，那么就存在一个与之对应的完美数，即$P=2^{p-1}(2^{p-1})$。读者朋友可以迅速验证，前四个梅森素数确实给出上面所说的前四个完美数。例如，用第三个素数5作为种子，我们得到完美数$P=2^4(2^5-1)=16\times 31=496$，即前述列表里第三个完美数。$P$的因数是直到$2^{p-1}$的2的各次幂，以及这些数乘上素数$2^p-1$。现在剩下要做的就只是一项练习了：将所谓的几何数列(geometric series，将在第5章中解释)求和，以便检查P的真因数之和确实是P。

在18世纪，伟大的瑞士数学家莱昂哈德·欧拉(Leohard Euler)进一步地证明了上述论断的逆命题，即每一个偶完美数都属于这一类型。这样，欧几里得和欧拉共同建立了一个梅森素数和偶完美数之间的一一对应关系。可是自然地，下一个问题出现了：所有的梅森数都是素数吗？很遗憾，并非如此。失败仅咫尺之遥，因为第五个梅森数等于$2^{11}-1=2047=23\times 89$。的确，我们甚至不知道梅森素数的数列是否会终结——也许最终，在某个点之后所有的梅森数都是合数。

尽管如此，梅森数依然是素数的天然候选，因为可以证明，一个梅森数m的任何真因数——假如存在的话——拥有$2kp+1$这样的特定形式。比如，当$p=11$，借助这个结论，我们只需检验被形如$22k+1$的素数除的情况。这两个素因数23和89，分别对应于值$k=1$和$k=4$。这个关于梅森数因数的事实还带来一个意外之喜，它提供第二种方法，使我们看出一定存在无穷多素数。因为它表明，2^p-1的最小素因数大于p，因而p不可能是最大的素数。由于这适用于任意素数p，我们可以推断不存在最大的素数，于是素数数列可以永远延续下去。

因为我们没有办法随心所欲地构造素数，所以在任一时刻，都存在着一个最大的已知素数。如今，这项桂冠总是落在一个梅森素数头上，这要归功于国际合作的互联网梅森素数大搜索项目(Great Internet Mersenne Prime Search， GIMPS)。这是一个始于1996年的志愿合作项目，它使用上千台并行工作的个人计算机，集成了一整套为此目的定制的算法来检验梅森数的素性。当前的世界之最是在2008年8月宣布的，它是 2^p-1, $p=43\ 112\ 609$ ，另外2009年4月又发现了一个新的p为42 643 801的梅森素数。这些数有大约1300万位，用普通的十进制记法需要上千页纸才能写下来[2]。

不那么完美的数

传统上人们对数的认识往往集中在单个数上，这些数被认为有特殊的甚至是奇妙的性质，就比如说完美数。不过，220和284是**一对**拥有类似特征的数。它们是第一对**相亲数**[3](amicable pair)，意思是每个数的真因数之和等于另一个——这是推广到数对的一种完美性。法国著名的业余数学家皮埃尔·德·费马(Pierre de Fermat)找到了其他的相亲数，如17 296和18 416，而欧拉更是发现了好几十对。出人意料的是，他们都漏掉了一对小数，1184和1210，这是由16岁的尼可罗·帕格尼尼(Nicol吘 Paganini)在1866年发现的。当然，我们还可以走得比数对更远一些，去寻找完美的三元数组、四元数组等。更长的循环比较罕见，但仍会出现。

我们可以从任何数出发，找到它真因数之和，接着继续重复这一过程，从而得到所谓的**真因数和数列**(aliquot sequence)。结果通常是令人失望的，因为我们一般会得到一条迅速抵达1的链，然后这个过程就终止了。举例说，即便是从一个看起来很有希望的数开始，比如12，链条还是很短：

$$12\to(1+2+3+4+6)=16\to(1+2+4+8)=15$$

→(1+3+5)=9→(1+3)=4→(1+2)=3→1.

困难在于，一旦你碰上一个素数，就结束了。完美数当然是例外，它们都会给我们一个循环，而相亲数则给我们一个双循环：220→284→220→…。

能产生超过两个元素的循环的数叫作**多亲的**(sociable)，相关研究直到20世纪才开始，因为那之前它们从没有被人发现过。甚至今天，还没有生成三元环的多亲数被找到，虽然现在已经知道了120个四元环数链。最早的一些例子是由P·普莱(P. Poulet)在1918年找到的。第一个是一个五元环数链：

12 496→14 288→15 472

→14 536→14 264→12 496.

普莱的第二个例子令人惊叹，时至今日还没发现其他能与之比肩的数链：从14 316开始，我们得到一个长度为28的循环。所有其他已知的循环长度均小于10。到今天，关于相亲数和多亲数，还没有像欧几里得和欧拉关于完美数那样漂亮的定理。不过，由于现代强大的计算能力，这类问题经历了一次由数值实验推动的复兴，人们也得出了一些新的结论。

根据一个数的真因数之和是小于、等于，还是大于这个数本身，我们可以将所有数划分为三类：**亏的**(deficient)、**完美的**(perfect)和**盈的**(abundant)。比如，就像我们已经看到的，12是一个盈数，18和24也是，因为它们的真因数之和分别为21和36。

在整数中进行初步的搜索，你可能由此猜测盈数也就是6的倍数而已。当然，任何大于6的形如6n的数都是盈的，因为6n的因数一定包含1， 2，3以及n， 2n， 3n，这些加起来大于原来的数6n。但是，这一

观察也可以被推广到不仅限于6的倍数，因为我们可以将同样的推理应用于任何完美数k。nk的因数将含有1，以及完美数k的所有因数乘上n所得出的数，于是nk的所有真因数加起来至少会得1+nk。所以，任何完美数的倍数都是盈的。例如，28是完美的，因而2× 28 = 56， 3× 28 = 84等都是盈的。

因此我们看到，完美数的倍数是盈数，同样的道理，盈数的倍数也一样。发现了这一点之后，你或许仍然会猜测，所有的盈数只是完美数的倍数。然而，你不用看太远，就会找到这个猜想的第一个例外，70是盈的，但它的因数没有一个是完美的。70是第一个所谓的奇异数(weird number)，不过不是因为上述原因(这个名字的源头下面会解释)。

有了这些发现，你可能还会认为，因为似乎不存在奇完美数，所以很可能也不存在奇盈数。换句话说，进一步的猜想可能变成了所有奇数都是亏的。倘若计算前几百个奇数的真因数和，似乎可以确证这一理论，但是这个说法最终还是被发现是错的。检验一下945，它的真因数和为975。现在，洪水的闸门被打开了，因为一个盈数的任意倍数都是盈的，所以945的奇数倍立刻给我们提供了无穷多奇盈数。

比起不假思索地逐个检验奇数，要是我们再精明一点的话，可能会更快地发现这个反例。要想一个数有很大的真因数和，它需要很多因数，其中还要包含大因数，这些大因数又是由小因数配对在一起产生的。于是我们可以通过将小素数乘起来**构造**具有大真因数和的数。如果我们只关心奇数，那么我们应该看由前几个素数——3，5，7等——构成的乘积。这个粗略的准则会很快使你检验到$3^3\times 5\times 7=945$，于是你也就在奇数中找到了盈性。

有时我们会发现，具有某些性质的数里，最小的也有很大的值，这种情况并不少见。尤其是当想找的数需要有某种因数结构，这种结

构是由你想要的性质决定的。于是那个最小的数可能极其大，不过如果我们在求解过程中利用给定的性质的话，它并不一定很难找到。这种数谜的一个例子是找到一个数，它既是一个立方数的5倍，又是一个五次方数的3倍。答案是

$$7\ 119\ 140\ 625=5\times 1125^3=3\times 75^5.$$

不过，并不难看出，为什么最小的答案都有数十亿这么大的值。任何解n都得是3^r5^sm这样的形式，r和s是正幂次，剩下的素因数被归总在整数m里，m不能被3或5整除。如果我们首先关注r的可能取值，可以观察到，由于n是**一个立方数**的5倍，指数(exponent)r一定是3的倍数。同时由于n是**一个五次方**的3倍，数$r-1$必为5的倍数。同时满足这两个条件的最小r是$r=6$。同样的，指数s一定是5的倍数，而$s-1$必须是3的倍数，最小的可行的s是$s = 10$。为了让n越小越好，我们取$m=1$，因此$n=3^6\times 5^{10}=3(3\times 5^2)^5=3\times 75^5$，于是$n$确实是一个五次方的3倍。同时$n=5(3^2\times 5^3)^3=5\times 1125^3$，所以$n$也是一个立方数的5倍。

一个更极端的例子是著名的牛群问题(Cattle Problem)，它是由古代最伟大的数学家阿基米德(Archimedes)提出的。但直到19世纪这个问题才被解决。要满足最初的44行诗中提出的所有限制条件，最小的牛群数量是一个超过200 000位的数!

上面这些讨论给了我们一条警示，那就是只有我们进入非常大的数的领域，数才会展示出它们全部的多样性。出于这个原因，仅仅是不存在少于300位的奇完美数这个事实，并不能说明它们“很可能”不存在。当然，假如真有一个出现了，这个领域内第一流的专家们也会大吃一惊。

让我们再次回到真因数和数列的一般行为这个话题上。我们还可以提出一些简单的问题，却仍然没有人回答得了。真因数和数列可能

的情形有哪些？如果这个数列遇上一个素数，那么在这之后它将立即终止于1，其实它也不会以任何其他方式终止。如果这没有终止，这个数列可能是循环的，从而代表了一组多亲数。但是，还有另外一种与之相关的可能性，我们可以通过计算95的真因数和来揭示：

$$95=5\times 19\rightarrow (1+5+19)=25$$

$$=5\times 5\rightarrow (1+5)=6\rightarrow 6\rightarrow 6\rightarrow \cdots.$$

在这个例子里，虽然95本身不是一个多亲数，但它的真因数和数列最终碰到一个多亲数(更准确地说是完美数6)，接着进入了一个循环。

可以设想，还存在一种可能性：一个数的真因数和数列永不遇见一个素数或多亲数。此时，这个数列必然是一个由不同数组成的无穷数列，其中没有一个是素的或多亲的。这可能吗？令人吃惊的是，没人知道。更惊人的是，存在一些小的数，它们的真因数和数列竟然还是未知的(因此它们可能拥有此类无穷真因数和数列)。这些谜一样的数中的第一个是276，它的真因数和数列由以下的数开始：

$$276\rightarrow 396\rightarrow 696\rightarrow 1104\rightarrow 1872\rightarrow 3770$$

$$\rightarrow 3790\rightarrow 3050\rightarrow 2716\rightarrow 2772\rightarrow \cdots.$$

但是没有人确切地知道它最后会变成什么样。

读者朋友可能会想自己探索一下，这样的话，让我来告诉你一个秘密，那就是如何从n的素因数分解来计算所谓的**真因数和函数**(aliquot function)$a(n)$：求所有可能的项$\frac{(p^{k+1}-1)}{(p-1)}$的乘积，这里$p$是素数，$p^k$是能整除$n$的$p$的最高次素数幂[4]，再减去$n$自己。例如，$276=2^2\times 3\times 23$，于是

$$a(276)=\frac{2^3-1}{2-1}\times\frac{3^2-1}{3-1}\times\frac{23^2-1}{23-1}-276$$

$$=7\times4\times24-276=672-276=396.$$

这与前面列出的276的真因数和数列的第二项相等。

对于与真因数和函数具有某种关系的数n，只需要通过给它们起个名字，我们便可以引入无穷多类的数。就像之前提到的，当a(n)=n时n是**完美的**，当a(n)＞n时n是**盈的**。一个**半完美数**(semiperfect number)n是等于自己部分真因数(小于n)之和的数，因而由定义可以推出，所有半完美数不是完美的就是盈的。比如，18是半完美的，因为18=3+6+9。当一个数是盈数但**不是**半完美的，它被称作**奇异的**(weird)，最小的奇异数是70。

你可能会认为，这个话题变得过于琐碎了——将名称冠予任意定义的数的类型，这举动确实不能让数字变得有意思，我们应该懂得在什么地方收手。话虽如此，要注意处理这些新问题背后的策略，还是欧几里得和欧拉展示给我们的关于完美数的那一套。回忆一下，欧几里得证明的是**如果**一个梅森数是素数，**那么**另一个数就是偶的且是完美的。欧拉则证明了反过程，所有偶完美数都是由这一途径产生的。在公元9世纪，波斯数学家塔比特·伊本·库拉(Thabit ibn Qurra)对于任意数n引入了一个三元数组，**如果数组里的数都是素数**，就可以构造一对相亲数。塔比特的构造方法在18世纪被欧拉进一步推广，但即便这个加强版公式似乎也只能产生一部分相亲数对，还有很多相亲数**不能**由这个构造产生。(现在已知差不多1200万对相亲数。)到了现

代，克拉维茨(Kravitz)给出了奇异数的一种基于素数的构造方法。这个公式成功找到了一个很大的有五十多位的奇异数。

本章和前一章是为了通过各种各样的实例，让读者朋友熟悉自然数的因数和因数分解。自然数也叫作**正整数**(positive integers)。这将帮助你为下一章做好准备，你将了解那些思想如何被应用于当代密码学(cryptography)——关于秘密的科学。

[1] 又称完全数或完备数。

[2] 2018年12月21日，已知的最大素数已更新为$2^{82\ 589\ 933}-1$。有兴趣的读者可以参阅GIMPS项目官方网站 https://www.mersenne.org/。

[3] 又称亲和数、友爱数、友好数。

[4] 素数幂即单一素数的正整数次方。

04 密码学：素数的秘密生活 Cryptography: the Secret Life of Pri mes

读者朋友现在应该能够体会，从很早开始，自然数的集合就已经被看作谜语和秘密的源泉，产出了很多直到今天都没能解决的问题。对我们中的不少人来说，这已经是继续进行关于数本身严肃研究足够的理由。不过其他人可能会有不同的态度：虽然这些难题可能耐人寻味、充满挑战，但是也可以想象，它们对人类文明的其他方面影响甚微。然而，这样的想法是个错误。

过去的几十年里，人们逐渐意识到，我们时不时会有保密的需求，某些普通信息构成的秘密可以被编码成关于数的秘密。现在，密码学已经得到了全面的应用。我们最为珍贵的秘密，无论是商业的、军事的、个人的、财务的、一般政治性的还是彻头彻尾丑闻性质的，都可以在互联网上被保护起来——用有关普通自然数的秘密。

化身为数的秘密

这一切都是怎么做到的呢？任何信息，无论是一首诗还是一份银行账单，一张武器设计图还是一套计算机程序，都可以用词语来描述。当然，我们可能需要拓展用来表示词语的字母表，使它不仅限于含有普通的字母。我们或许会加上数字符号和标点符号，包括代表词语之间空格的特殊符号。即便如此，我们希望传输的所有信息(包括生成相片和图表的指令)总可以由一张字母表来表达。让我们假设这张表包含的符号不超过一千个。我们可以数一数这些符号，然后用一个独一无二的数来表示每个符号。因为数的成本低廉，取之不尽。为了我们的目的，或许使用位数相同的数会比较方便。比如，每个符号都被一个独一无二的4位个体识别码(personal identification number，PIN)表示。我们可以将这些符号按顺序串连起来，从而得到一个很大

很长的数，里面包含故事的全部。我们要是愿意，甚至可以在二进制下做这件事。这样我们可以设计一个方法将信息转译成一长串0和1。于是，我们想要发送的每条信息都可以编码为一个二进制字符串，然后在接收端由具有相应程序的计算机解码，再被编译为我们都可以理解的普通语言。这就是我们的第一层领悟：要传递信息，从理论和实用两方面来说，能从一个人向另一个发送数字就足够了。

但是将信息变成数并不是关键的思想。明确一点说，将所有信息数字化的具体过程可以被藏起来，不被大众知晓，但这并不能形成有效的保护、免遭窃听。的确，从密码学的观点来看，我们可以将任何信息——所谓的**明文**(plaintext)——与代表它的数等同看待，于是便可以把数看作明文本身。这是因为我们假定，任何人都有办法掌握这两者互相转换的途径。只有当我们用别的数掩盖明文数码的时候，信息才真正具有了保密性。

密码学便是关于密码(ciphers)(机密的代码)的学问。让我向你介绍一些虚拟角色吧，他们经常出现在密码学所考虑的各类情境中。我们设想爱丽丝(Alice)和鲍勃(Bob)想互相通信，但不想让窃听者伊芙(Eve)听见[1]。我们也许会本能地同情爱丽丝和鲍勃，而将伊芙想象成坏人。但是这可能与真相相反，伊芙或许代表了正义的警方，努力保护着我们免受鲍勃和爱丽丝的邪恶计划的伤害。

无论参与者的道德立场如何，爱丽丝和鲍勃都可以运用一个古老的方法，来将伊芙排除在对话内容之外，哪怕伊芙截取了他们之间传递的信息。方法就是用密钥(cipher key)来加密数据，而这个密钥只有爱丽丝和鲍勃自己知道。他们可以预定在一个安全的环境会面，交换一个秘密数字(比方说57)，然后各自回家。当时机到来时，爱丽丝想要发送一条信息给鲍勃。这里为了叙述方便，我们假设信息可以用一个1～9之间的数字来表示。在那个重大的日子，爱丽丝想将信息

“8”发给鲍勃。她取出信息，加入秘密数字。也就是说，她通过加上57把真实的数值掩藏了起来，然后将信息8+57 = 65通过一个未经保护的渠道发给鲍勃。鲍勃收到这条信息，减去秘密数字，从而获取爱丽丝的明文65-57 = 8。

不过，不怀好意的伊芙很清楚这两个人在干些什么，并且她也确实截获了加密过的信息65。但是她能对它做什么呢？她可能像我们一样，知道爱丽丝发送给鲍勃九种可能的信息1，2，3，…，9中的一个，也知道她通过将信息加上一个数进行了加密，这个加密数一定在65-9 = 56到65-1 = 64之间。然而，因为她不知道这九个数字中的哪一个被使用了(她被排除在这个秘密之外)，她没法进一步搞清楚爱丽丝给鲍勃发送的明文信息。九种信息中的每一种都有相同的可能性。她知道的全部就是爱丽丝给鲍勃发送了一条信息，但是不知道信息的内容。

看起来爱丽丝和鲍勃对伊芙的邪恶渗透已经防得滴水不漏，他们似乎可以使用这个奇妙的数57来掩藏所有想说的内容，从而不受限制地通信。但是，实际却非如此。他们最好换一换这个数，实际上每次都使用一个新的密码会更好。不然这套系统就会开始泄露线索给伊芙。比如说，在未来某一周爱丽丝想向鲍勃传送一条相同的信息8。每件事都会与之前一样，再一次伊芙从电波中截获了神秘的数65，但这次会让她读出一些东西。伊芙会知道，无论这条信息是什么，这和爱丽丝第一周发给鲍勃的是一样的——而这类事情正是爱丽丝和鲍勃所不希望伊芙知道的。

当然，对于爱丽丝和鲍勃来说，这看起来也不是什么大问题。当他们第一次见面“互换密钥”时，爱丽丝可以不只是约定一个密码，而是提供给鲍勃一张长长的有顺序的列表，里面有上千个密码。他们

可以依次使用这些密码，从而避免了在公开通信中出现有意义重复的可能性。

我们确实也是这样操作的。这种密码系统在业内被称为**一次性密码本**(one-time pad)。发送者和接收者从密码本中找到一个一次性的数，用以掩藏他们的明文。当信息被发送和解密后，密码本中的那一页会被发送者和接收者一起撕掉。一次性密码本代表了一种彻底安全的系统，因为在公共领域传输的不安全文本不含有任何关于明文内容的信息。为了解码，拦截者需要拿到密码本才能获得加密/解密的钥匙。

密钥和密钥交换

一次性密码本似乎完全解决了安全通信的问题。某种角度上说确实如此。但是，参与者必须交换一份密钥，才能使用一次性密码本，类似的密码都有这样的问题。在实践中，交换密钥很麻烦。对于高层通信，比如在白宫和克里姆林宫之间，钱不是问题，所以必要的信息交流会在最高安全级别下开展。另一方面，在日常生活中，各类人和机构都需要以保密的方式互相沟通。参与者负担不起安全交换密钥所需的时间和精力，即使通过受信任的第三方来安排，加密可能依然价格不菲。

所有使用了几千年的密码——直到20世纪70年代——都有一个共同的缺点：它们都是**对称密码**(symmetric ciphers)，就是说加密和解密的钥匙在本质上是一样的。无论是尤利乌斯·恺撒(Julius Caesar)的简单的字母位移式密码，还是第二次世界大战中的复杂的英格玛密码(Enigma Cipher)，它们都摆脱不掉共同的弱点，即一旦对手知悉你是如何编码信息的，他们就能与你一样顺利地解码。为了使用对称密码，通信双方需要一种安全的密钥交换方式。

一直以来，我们似乎默认加密代码有不可避免的原则——要运用一套密码，双方需要用某种方式交换相应的密钥，并对敌人保密。的确，有人可能把它当作了数学常识。

这种假设恰恰会令一个数学家心生疑惑。本质上，我们面对的是一个数学问题，因而人们会预期这样一条“原理”有着坚实的基础，甚至表达成某种形式的数学定理。然而，并没有这种定理。之所以没有，是因为这条原则根本就是不对的。下面这个思想实验可以证明。

从爱丽丝那里传送一条机密信息给鲍勃，这本身并不一定需要交换一份密钥，因为他们可以按下面的程序来操作。爱丽丝写好她要给鲍勃的明文信息，然后将它放进一个盒子里，再挂上她自己的锁，只有爱丽丝有锁的钥匙。她接着将盒子寄给了鲍勃。当然，鲍勃没法打开它，不过他可以给盒子加上第二把锁，而只有他一个人有这把锁的钥匙。接下来，盒子被寄还给爱丽丝，她会打开自己的锁，第二次把盒子寄给鲍勃。这次，鲍勃便可以打开盒子，读到爱丽丝的信息。寄送过程中，我们知道伊芙即使想从中作梗，也没办法看到内容，因此信息是安全的。这样，一条机密信息便可以通过不安全的渠道安全地传送，却从不需要爱丽丝和鲍勃交换密钥。这个假想的情境说明，在密信的传递过程中，没有定律规定密钥**必须**转手。在真实系统中，爱丽丝和鲍勃的“锁”可能是他们各自对信息的编码，而不是一个分隔潜在窃听者和明文的实体设备。爱丽丝和鲍勃可以利用这个初始的交换来建立一套普通的对称密码，接下来这套密码便可以掩藏未来的所有通信。

其实在真实世界中，安全的通信渠道经常是这样建立起来的。不过，用个人代码代替实体的锁并不是那么容易。解读(即解锁)的过程需要爱丽丝在先，鲍勃在后。不像普通的锁，爱丽丝和鲍勃的编码可能会相互干扰，从而导致这一过程无法进行。但是，1976年，惠特菲

尔德·迪菲(Whitfield Diffie)和马丁·赫尔曼(Martin Hellman)首次公开证明了这个方法是有效的。

另一种相关的方法是**不对称**(asymmetric)或者说**公开密钥加密**(public key cryptography)的思想。在这个方法中，每个人发布他们自己的公开密钥，要发送给一个人的信息都用他的公开密钥来加密。但是，每个人还有一份私人密钥。如果没有私人密钥，使用对应的公开密钥加密的信息就没法解读。如果接着用挂锁比喻的话，爱丽丝提供给鲍勃一个盒子用来装他的明文信息，外加一只开着的锁(她的公开密钥)，而只有她自己手里握着钥匙(她的私人密钥)。

看起来，建立一个实用的公开密钥系统需要满足太多条件，因为安全性和易用性这两个要求密不可分，但似乎又互相矛盾。不过，快速、安全的加密方法已经在互联网上广为应用了，即便大家很少意识到它的存在和它为人们提供的保护。让这一切得以实现的，说到底都是数，尤其是素数。

用素数的秘密守护我们的秘密

回忆一下，我们把每条明文信息都看作一个单独的数，我们自然努力想用其他数来掩盖它。最常用的方法是用所谓的RSA(Rivest-Shamir-Adleman)加密过程，这是由它的创始人罗纳德·李维斯特(Ron Rivest)、阿迪·萨莫尔(Adi Shamir)和莱昂纳德·艾德曼(Leonard Adleman)在1978年发表的。在RSA中，每个人的私有密钥由三个数p，q，d组成，p和q是(非常大的)素数，而第三个数d是爱丽丝保密的解密数，我们到下面合适的时候再解释它。爱丽丝公开给大家的是两个秘密素数之积$n=pq$，以及一个加密数e(这是一个普通的整数，与第2章中提到的特殊常数e没有任何关系)。

为了说明RSA如何发挥作用，我们举一个简单的例子。假设爱丽丝的素数是$p= 5$和$q= 13$，于是$n = 5\times 13 = 65$。如果爱丽丝把她的加密数设为$e = 11$，那么她的公开密钥就是$(n, e) = (65, 11)$。为了加密信息m，鲍勃只需要n和e。然而，要想解码鲍勃发送给爱丽丝的经过加密的信息$E(m)$，就需要有解码数d。在这个例子中，可以推知$d= 35$，这个我们待会儿就来证明。计算出d的数学方法需要知道素数p和q。在这个简单的例子里面，给定$n = 65$，任何人都会很快发现$p = 5$而$q = 13$。然而，如果素数p和q**极其**地大(通常它们有几百位数字那么长)，在实际操作上，至少是在较短的时间(比如说两到三周内)，几乎没有任何计算机系统能完成这项任务。总的来说，RSA加密系统是基于一个经验事实，即找到一个很大很大的数n的素因数极为困难，困难到在实际操作中无法实现。这个方法设计得真正聪明的地方，在于代表信息的数m可以只用公开的数n和e来加密，但在实际操作中，解密需要知晓n的素因数。在这章剩下的部分里，我们就来详细解释这一点。

RSA是这样起作用的：鲍勃通过网络发送的不是m本身，而是m^e除以n所得的**余数**(remainder)。然后爱丽丝拿到这个余数r，计算r^d除以n得到的余数，就能重新得到m。这背后的数学保证了爱丽丝得到的结果正是原始的信息m，爱丽丝的计算机可以进一步将它解码为普通明文。当然，对于任何真实生活中的“爱丽丝”和“鲍勃”，这一过程都是在幕后无缝衔接地完成的。

这么看来伊芙所缺的唯一重要的东西是解密数d。倘若伊芙知道d，那她也能像爱丽丝一样完美地解码信息。事实上，存在着一个特殊的方程，d恰好是它的一个解。从计算的角度来讲，借助于公元前300年的《几何原本》[2](Books of Euclid)中发表的欧几里得算法，求解这个方程是很容易的。这不是困难之处。麻烦的地方在于，除非你知道p和q中至少一个，否则不可能准确地找到那个要解的方程。这便是挡在伊芙道路上的障碍。

我们可以再进一步解释，以上涉及的数如何在这个系统中各司其职。首先，很明显鲍勃一开始就面临着一个大问题，m很大，n更是可怕(在200位数字这个量级上)。即便e的值不像那样夸张，m^e也是极其大的。计算出它之后，我们还得将m^e除以n来得到余数r，这代表了被加密的文本。这些计算太过繁杂，以至于看起来也许并不可行。我们需要意识到，即使现代计算机异常强大，它们还是有能力极限。当计算涉及很高次的幂，就可能会超过**任何**计算机处理能力的极限。我们不能假设电脑可以在短时间内完成任何我们交给它们的计算任务。

鲍勃有根救命稻草是，他完全可以不做很长的除法就找到要求的余数r。实际上，余数仅仅取决于余数。这里我们举一个例子来说明，7^{39}的最后两位是多少？(换句话说，当这个数被100除后余多少？)为了回答这个问题，我们可以从计算7的前几次幂开始：$7^1 = 7$， $7^2 = 49$， $7^3 = 343$， $7^4 = 2401$，$7^5 = 16\ 807$，…。不过很快我们就清楚地发现，离7^{39}还远着的时候，这些数已经变得相当大，我们都处理不过来了。另一方面，随着我们算出一个又一个幂次，一个关键的规律出现了。当计算连续的幂次时，结果的最后两位数字仅依赖于前一个数的最后两位。这是因为我们做乘法时，百位及以上的数字并不会影响到结果在个位和十位上的数字。

同时，因为7^4末尾两位是01，所以接下来四个幂的末尾依次是07， 49， 43，接着又是01。因此，随着我们挨个计算幂次，末尾两位的数字只会一遍又一遍地重复这个长度为4的循环。回到我们手上的问题，由于$39 = 4\times 9+3$，我们会经历这个循环九次，然后还需要三步来计算7^{39}的最后两位数，因此结果一定是43。

这样的技巧是相当普适的。比方说，为了找到某个幂次a^b除以n所得的余数，我们只需要取a除以n所得的余数r，并追踪r的各次幂除以n之后的余数。余数r一定是大于或等于0、小于或等于$n-1$之间的一个

数。在我们只关心r的时候，数学家们说我们是在求**模**(modulo)n。我们舍弃了所有n的整数倍，因为它们除以n余0，所以不会对最终的余数r有任何贡献。

你可能还是在怀疑，我是不是通过选择特殊的例子操纵了证据，这个例子里一个很小的幂次就给出了余数1——这里7^4比100的某个整数倍大1。然而，你怀疑的情况只在部分程度上成立。事实上，如果我们取任意两个数，a和n，它们的最大公因数为1，我们说这些数是**互素的**(coprime)，那么总存在一个指数t，使得幂a^t等于1模n——也就是说被n除时余1。从这个角度说，连续次幂的余数会构成周期为t的循环。但是，预测t的值却很难。不过人们知道t总等于一个特殊的数或是它的某个因数。这个数传统上被写作$\phi(n)$，代表欧拉ϕ **函数**(phi function)的值。

那么什么是$\phi(n)$？它被定义为小于n的与n互素的数的个数。例如，设$n=15$，那么我们关心的数集为{1，2，4，7，8，11，13，14}，这里我们列出了所有直到15、但与15没有共同的因数的数(除了不可避免的因数1)。因为这个集合有8个元素，所以可以看出$\phi(15)=8$。幸运的是，有一个更简洁的方法可以找到$\phi(n)$，不需要总是明确地列出所有小于n并与之互素的整数，再数它们的个数。跟大部分有类似特性的函数一样，这个值可以用n的素因数分解来表达。对，我们只需要知道n的素因数，因为用n乘以所有可能的分数$(1-\frac{1}{p})$就可以得到$\phi(n)$，这里p遍历n的所有素因数。比如，15的素因数为3和5，因此这里答案是

$$\varphi(15)=15\times(1-\frac{1}{3})(1-\frac{1}{5})=15\times\frac{2}{3}\times\frac{4}{5}=8\ .$$

这跟我们直接从定义得到的结果是一样的。使用这个方法，你可以自己验证$\phi(100)=40$。因此，可以推出7^{40}同余于1模100。当然，就像我们已经看到的，余1的7的最小幂次不是40，而是它的因数4。

以上这些都是为了表明，鲍勃确实可以求出发送给爱丽丝的数，m^e模n，同时不需要鲍勃的计算机做出太多努力。不过，在实际操作中涉及的数依然大到可怕，因此我们有必要进一步说明如何处理它们。计算m^e所涉及的高次幂可以分阶段进行，这个过程叫作**快速求幂**(fast exponentiation)。简而言之，这一方法运用连续求平方以及幂的相乘来得出m^e模n，我们可以用二进制形式的e引导算法，从而在相对少的步数里快速找到想要的余数。

欧几里得教爱丽丝找到她的解密数

解密数是接收者的魔杖，能让信息重见天日。挑选数d的标准是使乘积de除以$\phi(n)$余1。因为$n=pq$是两个不同的素数的积，所以函数值$\phi(n)=pq(1-\frac{1}{p})(1-\frac{1}{q})=(p-1)(q-1)$。其实，在小于$\phi(n)$的范围中，具有想要的性质的数$d$有且只有一个。

为了找到d，爱丽丝的电脑可以利用一个代数工具——欧几里得算法，这一工具已经有2300多岁了。稍后我们就来介绍它。如果伊芙的计算机知道去解哪个方程，那它当然也能做到同样的事情。然而，因为p和q是爱丽丝私有的，$(p-1)(q-1)$也是，因此伊芙并不知道从哪里着手。

加密数(公开的)e需要满足一个温和的限制条件，才能保证d的存在。爱丽丝必须确保e与$\phi(n)$没有公因数。这很容易做到，爱丽丝可以检验用不同的素数除$\phi(n)$的结果，从而保证既不泄露p和q的值也让e满足限制条件。实际应用中，e的值常常使用第四个所谓的**费马素数**(Fermat prime)$e=65\ 537=2^{16}+1$。这个值，$2^{24}+1$，具有一个尤其稀有

的性质，即可以用直尺和圆规(straightedge and compass)作出一个有e条边的正多边形。不过，它在密码学中的用处在于它是一个很大的(恰好比某个2的幂大1的)素数，这非常适合运用之前提到的快速求幂过程。

回到欧几里得算法。这个算法是通过推广以下的观察结果得到的，即对于两个数$a>b$，可以通过连续的相减来找到**最大公因数**(highest common factor)，最大公因数也叫作**最大公约数**(greatest common divisor)。我们注意到$r = a-b$有一项性质：a，b，r中任意两者的公因数也是第三个数的因数。例如，如果c是a和b的公因数，于是$a = ca_1$并且$b = cb_1$，我们看到$r=a-b=ca_1-cb_1=c(a_1-b_1)$，这就得出了$r$包含$c$的一个因数分解。于是，$a$和$b$的最大公因数等于$b$和$r$的最大公因数。因为这两个数都小于$a$，我们现在面对的是跟之前一样的问题，不过是相对于两个更小的数。重复应用这个方法，最终会得到一对数，它们的最大公因数可以一眼看出。(实际上，最后手中的两个数会变成相等的，因为如果不是这样，我们可以继续多做一步。这对数共同的值就是我们找的那个数。)

比如，要找$a = 558$和$b = 396$的最大公因数，第一次减法后我们得到$r = 558-396 = 162$，因此新数对是396和162。由于396-162 = 234，所以我们的第三对数是234和162。随着我们继续下去，完整的数对依次是：

$$(558, 396)\to(396, 162)\to(234, 162)\to(162, 72)\to(90, 72)\to(72, 18)\to(54, 18)\to(36, 18)\to(18, 18).$$

因此558和396的最大公因数是18。

还可以根据待考察的数对各自的素因数分解来写出它们的最大公因数。在这个例子里，$558=2\times 3^2\times 31$，而$396=2^2\times 3^2\times 11$；取两个分

解中各个素数的共同次幂，我们得最大公因数为$2\times 3^2=18$。然而，对于大数而言，使用欧几里得算法需要的工作量少得多，因为一般执行减法运算比寻找素因数分解更简单。

欧几里得算法的另一项优点是总可以倒着做，这样便可以将最大公因数用原始的两个数来表示。为了更好地看清楚这是怎么做到的，我们在刚才这个例子里将计算压缩。在依次相减的过程中，同样的数重复出现了好几次，我们可以把它们表达在一个方程里。这样我们就得到以下几个等式：

$$558 = 396+162;$$

$$396 = 2\times 162+72;$$

$$162 = 2\times 72+18;$$

$$72 = 4\times 18 .$$

从第二行开始到最后一行，我们可以用各个等式逐个消去中间余数。在这个例子中，先利用倒数

第二个等式，然后是它上面的那个，我们得到：

$$18=162-2\times 72=162-2\times (396-2\times 162)$$

$$=5\times 162-2\times 396.$$

最后再用第一个等式，我们可以把第一个中间余数162也消去：

$$5\times 162-2\times 396=5\times (558-396)-2\times 396=$$

$$=5\times 558-7\times 396=18.$$

无论是对于实践还是理论，可以倒过来执行这一程序都是很重要的。特别是在解密里，为了找出爱丽丝的解密数d，我们想要d满足de

除以$\phi(n)$余1的条件。为了简洁，我们用一个单独的符号k来代表$\phi(n)$。现在就可以看出我们坚持要e和k互素的原因了。因为如果它们的最大公因数是1，当我们对e和k执行欧几里得算法时，最终出现的余数自然是1。倒过来执行这一算法，我们就能最终把1表示成e和k的组合。特别地，我们会找到整数c和d，它们满足$ck+de = 1$，或者换句话说$de=1-ck$。因此de除以k会余1。

这个相对简单的过程将给出爱丽丝的解密数d：直接从方程里得出的d的值可能不在1到k的范围内；倘若不在，通过加上或减去合适的k的倍数，我们最终可以找到那个d，在给定范围内，它是唯一具有de除以k余1这个神奇性质的数。(我们可以轻松证明d的唯一性，不过这里还是不要离题太远了。)这便是如何计算解密数d的方法。我们可以回到前面的例子来说明，这里$p=5$，$q=13$，于是$n=pg=5\times 13=65$。我们有$\phi(n)=(p-1)(q-1)=4\times 12=48$。爱丽丝设定$e=11$，由于11与48互素，这在游戏规则所允许的范围内。应用欧几里得算法于$\phi(n)=k=48$和11，可得：

$$48 = 4\times 11+4;$$

$$11 = 2\times 4+3;$$

$$4 = 1\times 3+1 .$$

这显示k和e的最大公因数确实是1。将算法反转，我们得到：

$$1=4-3=4-(11-2\times 4)=3\times 4-11$$

$$=3(48-4\times 11)-11=3\times 48-13\times 11.$$

为了满足$11d$除以48余1的条件，我们解出d的初步值-13。为了得到一个在要求范围内的正的d的值，我们在这个数的基础上加上48，于是$d=48-13=35$。

d能帮助爱丽丝解密信息，其原因可以归结为模算术(modular arithmetic)的特性，以及当de除以$k=\phi(n)$时余1这一事实。爱丽丝计算$(m^e)^d=m^{de}$模n。现在de具有$1+kr$的形式，这里r是某个整数。正如之前解释的那样，m^k除以n余1，这通常被称为**欧拉定理**(Euler’s Theorem)，而这对于$(m^k)^r=m^{kr}$也是对的。因此，$m^{1+kr}=m\times m^{kr}$，它除以n余m。(详细的验证需要一点代数运算，不过结果的确是这样的。)通过这个方法，爱丽丝得到了鲍勃的信息——m。

顺便指出，我们在证明素因数分解唯一性的时候缺少了一环，欧几里得算法恰好提供了这缺失的环节。因为它使得我们能够验证欧几里得性质，即如果素数p是乘积ab的一个因数，比方说$ab=pc$，那么p是a和b之中至少一个的因数。如果p不是a的一个因数，那么由于p是素数，a和p的最大公因数是1。应用欧几里得算法于a和p这对数，并将其反转，我们可以找到整数r和s，使得$ra+sp=1$。这已经足够证明p是b的一个因数了，由于$ab=pc$，我们有：

$$b=b\times 1=b(ra+sp)=r(ab)+psb:$$

$$=r(pc)+psb=p(rc+sb).$$

这便是想要找的b的分解，它显示p正是一个因数。

总而言之，RSA加密背后的关于数的理论保证了系统的可靠性。当然，要保证系统的完整，还需要遵守很多这里没有解释的协议。可能出现的问题包括**身份验证**(authentication)(假如伊芙伪装成鲍勃联系爱丽丝怎么办？)、**不可抵赖性**(nonrepudiation)(假如鲍勃装作伊芙发送了信息给爱丽丝怎么办？)以及**身份欺诈**(identity fraud)(假如爱丽丝滥用鲍勃发给她的保密的身份信息，试图在网上假扮他怎么办？)。另外，当可预测的或者重复的信息大量出现，这个系统的其他弱点也可能会暴露。不过，这些困难在任何公开密钥加密系统中都可

能出现。它们是可以被克服的，并且总体来说与背后的数学技术没有太大关系，而那些数学技术才是保证加密的高质量和稳定性的因素。

这一章展示了素数以及整除性和余数理论的一项主要应用。我们不仅从广义的原理上，还在细节层面对此进行了解释，这都要感谢欧几里得的古代数学和欧拉在18世纪的贡献。

我们这本书的第一部分会在第5章告一段落，该章里我们将要介绍一些特殊类型的整数，它们与某些自然呈现的分组现象有关。

[1] “窃听者”的英语单词为eavesdropper，因此虚拟的窃听者常用名字Eve来代表。

[2] 一般称为The Elements of Euclid，共有13卷。

05 计数的数 Numbers That Count

从计数问题中自然产生的数很重要，因此它们已经被研究得很深入了。这里我将介绍二项式系数，以及卡特兰、斐波那契和斯特林所发现的数。这些数被用于枚举某些自然形成的集合。不过，还是让我们从一些非常基本的数列开始吧。

三角形数、算术数列和几何数列

因为它们在二项式系数里还会出现，让我们花点时间复习一下三**角形数**(triangular numbers)。这一数列的第n项，记为t_n，定义为前n个正整数的和。它的值依赖于n，可以用下面这个技巧来计算。我们将刚刚提到的和的形式写作t_n，接着以倒序再写一遍：

$$t_n=1+2+3+\cdots+(n-2)+(n-1)+n.$$

$$t_n=n+(n-1)+(n-2)+\cdots+3+2+1.$$

将这两个版本的t_n相加，结果显而易见是$2t_n$。但我们这么做是为了将1和n、2和$n-1$、3和$n-2$配成对。每一对的和都相等，为$n+1$，且一共有n个这样的数对。于是可以推得$2t_n=n(n+1)$。换句话说，第n个三角形数是$\frac{1}{2}n(n+1)$。例如，从1到1000所有整数之和为500× 1001 = 500 500。

事实上，用这个方法我们可以得到对形如a，$a+b$，$a+2b$，$a+3b$，…的**算术数列**的前n项求和的一般规则。我们首先处理比例上的变化。将t_n的表达式乘以b，我们得到$b+2b+3b\cdots+nb=\frac{b}{2}n(n+1)$。然后计算它的前$n-1$项的和——只需在刚刚得到的公式中用$n-1$代替$n$，即$\frac{b}{2}n(n-1)$。现在，再给每一项加上$a$，包括作为首项的额外的$a$，就得到了最一般的算术数列。这

意味着我们需要向刚才的和再加上na。于是，算术数列前n项的一般求和公式为：

$$a+(a+b)+(a+2b)+\cdots+(a+(n-1)b)=na+\frac{b}{2}n(n-1).$$

例如，取a=1和b=2，则前n个奇数之和为$n+n(n-1)=n+n^2-n=n^2$，即n的平方。

如果把加法操作替换为乘法，算术数列就变成了**几何数列**[1](geometric series)。算术数列中，相邻两项相差一个**公差**(common difference)，即我们的b。换句话说，从一项到下一项，我们要加上b。几何数列中，我们还是取一个任意数a为首项，但是通过**乘**一个固定的数——称为**公比**(common ratio)，来得到下一项。这个比值记为r。也就是说，典型的几何数列具有a，ar，ar^2，…的形式，其第n项为ar^{n-1}。就像算术数列一样，几何数列的前n项和也有一个公式[2]：

$$a+ar+ar^2+\cdots+ar^{n-1}=\frac{a(r^n-1)}{r-1}.$$

要想看出这一公式的正确性，最快的方法是将等式两边同乘(r-1)并将括号展开。等式左端我们有：

$$(ar+ar^2+ar^3+\cdots+ar^n)-(a+ar+ar^2+\cdots+ar^{n-1}).$$

这一表达式可以**裂项相消**(telescope)，即一个括号中，几乎每一项都可以与另一括号中的一项互相消去，仅剩下$ar^n-a=a(r^n-1)$。由

此可见我们的求和公式是正确的。举个例子，设$a=1$，$r=2$，我们得到2的各次幂之和：

$$1+2+4+\cdots+2^{n-1}=2^n-1.$$

第3章中欧几里得通过梅森素数导出了偶完全数，这里的公式恰好可以让你验证他的方法。

阶乘、排列和二项式系数

我们已经看到，第n个三角形数来自于对1到n之间所有的数求和。这个过程中，倘若将加法换成乘法，我们就得到了所谓的**阶乘**。这一概念已经在第2章中出现过。

阶乘常常现身于计数和概率问题中，比如打扑克时抽到特定牌的概率。因而它有单独的记号：n的阶乘记为$n!=n\times(n-1)\times\cdots\times 2\times 1$。随着$n$的增长，三角形数的大小增长得相当快，增长率差不多能达到n^2的一半，然而阶乘变大还要快得多——很快就能增大到百万量级。比如10! = 3 628 800。阶乘由感叹号来代表，恰恰提醒着我们那令人惊叹的增长速度!

具体来说，$n!$表示了排列n个物体有多少种不同的方法，如将n个球排成一串 。这是因为第一个位置你有n种选择；确定第一个球之后，第二个位置剩下$n-1$种选择；第三个位置有$n-2$种选择，以此类推。如果我们仅限于选择r个球做排列，并将不同方法的数目记为$P(n, r)$，我们说从n个球中一次取出r个球一共有$n\times(n-1)\times(n-2)\times\cdots\times(n-r+1)$种**排列**(permutation)。这也可以方便地用两个阶乘的比值来表示：$P(n, r)=\frac{n!}{(n-r)!}$. 例如：你从52张扑克牌中抽取5张牌，一共存在$P(52, 5)=52\times 51\times 50\times 49\times 48$种可能的结果。当然，打牌时所谓的手牌与抽到的顺序无关，只要是这些牌就行了。这5张牌有5! = 120种不同的排列

方式。因此，一套5张手牌真正不同的组合有$\frac{P(52, 5)}{120}=2\ 598\ 960$种，大约有250万之多。

在关于计数的问题——或称为**枚举**(enumerations)问题中，最特殊的一类数便是**二项式系数**(binomial coefficients)。之所以叫这个名字，是因为它们是将二项式$(1+x)^n$展开后x的各阶幂次前的系数。二项式系数$C(n, r)$代表了我们从n个元素中挑选出r个元素一共有多少种不同的方法。例如，$C(4, 2)=6$，因为从4个元素中取一对有6种方法(不在乎这两者的次序)。再具体些，假设我们有4位儿童：A，B，C和D。那么有6种方法从中选出1对：AB，AC，AD，BC，BD和CD。

二项式系数可以用两种不同的方法计算。第一，我们可以推广之前我们用来计算$C(52, 5)$的办法，我们发现$C(n, r)=\frac{P(n, r)}{r!}$，于是我们得到以下实用的公式：

$$C(n, r)=\frac{n!}{(n-r)!r!}.$$

一种值得注意的情形是当$r=2$时，即从n个物体中选出一对的方法个数。答案是$\frac{n!}{(n-2)!2!}$。分母中$(n-2)!$项的所有因数都与中$n!$相同的因数消去了，同时$2! = 2$，表达式$C(n, 2)$因而简化为$\frac{n(n-1)}{2}$。换句话说，从n个物体中选出一对共有t_{n-1}种不同方法——第$n-1$个三角形数。比如，我们已知道$C(4, 2)=6$，确实是第3个三角形数。

这个用阶乘计算二项式系数的公式确实提供了一种简便的代数方法，使我们能够证明这些系数的许多特殊性质。然而，如果我们用第二种方法来导出这些整数，这些性质的演化会更清楚。这种方法基于**算术三角形**[3](Arithmetic triangle)(如图2)，又称为**帕斯卡三角形**(Pascal’s triangle)，以纪念17世纪法国数学家和哲学家布莱士·

帕斯卡(Blaise Pascal)。算术三角形在过去的1000年里被波斯、印度和中国数学家各自发现。它出现在1303年朱世杰[4]所著《 四元玉鉴》的封面上。

这个三角形内部的每个数都是其上方两数之和，它给出了所有二项式系数，可以无限续写下去。比方说，要找出从7个人中选出5个共有多少种方法，可以遵循以下步骤：将三角形各行编号，从上向下由0开始。在每行中将元素从左向右编号，依然从0开始。取行编号为7列编号为5的数(记住从0开始数)，答案是21。你注意到每行有对称性：如21也是从7人中挑出2人的方法总数。当从7人中选出5人时，我们也同时挑出了2人，即剩下的那2位。这就解释了上面的现象。这一对称性原则自然适用于每一行。它还体现在公式$C(n,r)=\frac{n!}{(n-r)!r!}$中：用n-r代替r，我们还是得到同样的公式，因为分母上仅仅是r和n-r调换了位置。

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
⋮

图 2　算术三角形

算术三角形的结构能够给出正确的答案，这一点不难理解。每一行都由上一行所产生。我们可以轻松看出前三行是正确的：例如，第三行中央的2告诉我们从一对人当中选出一位总共有两种方法。顶端的1表明从空集中选出0个元素就只有一种方法。实际上，从任意集合中

选出0个元素的方法都是一种，这就是为什么每行都从1开始。我们重点看刚才的例子——共有21 = 15+6种方法从7人中选出5人。这21种选法自然地分成两类。第一类，有15种方法从前6人中选出4人，我们可以再加上第七位凑成5人组。或者如果我们不选第七位，则要从前6人中一次选出5人，共有6种方法。这个例子告诉我们一行是如何生成下一行的：每个元素都是上面两数之和，按照这一模式从上到下建立起整个三角形。这个规则可用符号表示成以下形式：

$$C(n, r) = C(n-1, r) + C(n-1, r-1).$$

算术三角形蕴含着丰富的规律。比如：将每行的数分别相加可以得到倍增数列1， 2， 4， 8， 16，32，…，即2的各次幂。以1，8， 28， 56，…这行为例，我们是在将从8个元素中选出0个、1个、2个、3个……元素的方法个数相加，最后得到的是从8个元素中一次选取**任意**个数的元素共有多少种方法。这一数字为2^8，因为一般一个有n个元素的集合拥有2^n个子集。

上面这一事实也可以直接推出，原因是一个n个元素集合的任意一个子集可以使用长度为n的二进制字符串来识别。方法如下：我们考虑一个集合，比如$\{a_1, a_2, \cdots, a_n\}$，则一个长度为$n$的二进制字符串定义了它的一个子集，因为字符串中的每个1表示相应的元素a_i存在于我们的子集中。例如，假设$n=4$，字符串0111和0000分别代表$\{a_2, a_3, a_4\}$和空集。由于二进制字符串的每个位置都有两种取值选择，因此共有2^n个这样的字符串，一个n元素集合便含有2^n个子集。

卡特兰数

算术三角形中，从第一行开始，每两行就会有一个数位于正中央：1，2，6，20，70，252，924，…。这些数分别被连续的正整数1，2，3，4，5，6，7，…整除。所得的一列商1，1，2，5，14，42，

132，…被称为**卡特兰数**[5](Catalan numbers)。用**中央二项式系数**表示的话，第n个卡特兰数可以写成$\frac{1}{n+1}C(2n，n)$，n在0，1，2，…中取值。

这种数有种最简单的图形表达，是用n段上斜线段和n段下斜线段能画出多少组不同的“山脉”(见图3)。

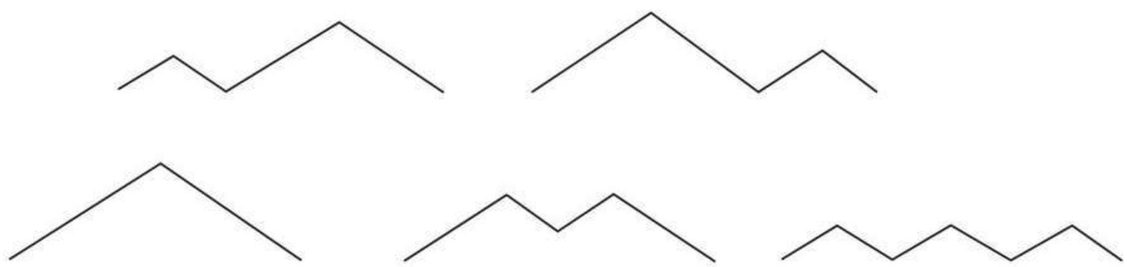


图3　用3段上斜线段和3段下斜线段，共有5种山脉形态

每种不同的山脉构型都对应一组有意义的括号，因此将n对括号有意义地排列起来的方法个数，恰好是第n个卡特兰数。例如，(())()和((()))是有意义的括号方法，但())(()不是：有意义是指从左向右数时，左括号的个数从不小于右括号的个数。这对应于山脉始终位于地面上方这一自然条件。比方说，图3中第一个和最后一个山脉的构型分别对应于()(())和()()()这两种括号排列。

第n个卡特兰数还代表将n+2边的正多边形被互不相交的对角线分成三角形的方法个数。沿着这一思路，卡特兰数还有其他解读方法。正如二项式系数，也有公式联系了卡特兰数和更小的卡特兰数，这使对它们的计算变得很简便。

斐波那契数列

恐怕没有第二个数列像斐波那契数列(Fibonacci sequence)那样使普罗大众着迷了，它是如下的数列：

1， 1， 2， 3， 5， 8， 13， 21， 34， 55， 89，144，233， 377， 610，…

在起始的两个数之后，每个数都是其之前两数之和。在这一点上，二项式系数与其有相似之处，因为那里每一项也是之前两数之和。但是斐波那契数列的组成方式更简单：

$$f_n=f_{n-1}+f_{n-2}$$

这里f_n表示第n个斐波那契数，并且我们规定$f_1=f_2=1$。我们将这种用先行项来定义当前项的公式叫作**递归**(recursion)或**递推关系**(recurrence relation)。

这一数列是从哪里来的呢？它最先是由比萨的莱昂纳多(Leonardo of Pisa)——更有名的称呼是斐波那契——在他著名的兔子问题中引入的。一只雌兔出生两个月后达到生育年龄，并在这之后的每个月生下一只雌兔。那么每个月初雌兔的总数由斐波那契数列给出。第一个和第二个月初当然只有一只兔子。第三个月初雌兔生下一只雌兔，因而我们有2只雌兔。到了下个月，它又生下一只，于是共有3只雌兔。再下个月，雌兔总数达到5只，因为雌兔和它的大女儿都能够生育。一般地，在这之后的每个月初，**新生**雌兔的数量等于两个月前雌兔的总数，因为此刻只有它们处于生育年龄。于是，每月初雌兔总数等于上月雌兔的数量与上上个月雌兔数量之和(斐波那契的雌兔是永生的)。因而斐波那契数列的产生方式完全符合他的雌兔繁殖的方式。

尽管真实世界的兔子并不是以这种异想天开的方式来繁殖的，斐波那契数列依然换着面孔出现在自然界中，包括植物的生长。我们对这一现象的原因已经有了透彻的理解，这与该数列的更微妙的性质有关，即**黄金分割比**(Golden Ratio)。我们这就来谈谈它。

最简单的数列类型便是我们在本章第一部分介绍的算术和几何数列。虽然斐波那契数列并非它们中的一种，它却与后者有惊人的联系。当计算斐波那契数列邻项之差并将它们也排成一列时，我们得到0， 1， 1， 2， 3， 5， 8， 13，…，于是又得到了一组斐波那契数列，只是这次是从0开始的! 其中的原因正是这个数列形成的方式：两个相邻斐波那契数之差恰好等于它俩之前的那个数，(想要代数地证明这一点，可以将上面的斐波那契递推公式两边同时减去f_{n-1})所以它不是算术数列。它也不是几何数列，因为相邻两项斐波那契数之比并不是常数。但我们观察邻项之比的时候，它似乎在一个极限值附近稳定下来，而且这个比值很快就趋于稳定了。让我们将每个斐波那契数除以它前一个数：

$$\frac{34}{21}=1.6190,\quad \frac{55}{34}=1.6176,\quad \frac{89}{55}=1.6182,\quad \frac{144}{80}=1.6180,\quad \cdots.$$

这个逐渐显露的神秘的数，1.6180…，到底是什么呢？这个数τ被称为**黄金分割比**，它也会出现在一些几何问题中，而这些问题看起来跟斐波那契的兔子相差十万八千里。例如，τ 是正五边形的对角线与边长之比(如图4)。任意两条对角线的交点将它们各自分成两段，这两段长度之比也是τ： 1。一对相交的边和一对相交的对角线构成一个菱形(一个“正方的”平行四边形)，即如图4中的ABCD。所有对角线的交点又组成了一个小的倒立的五边形。

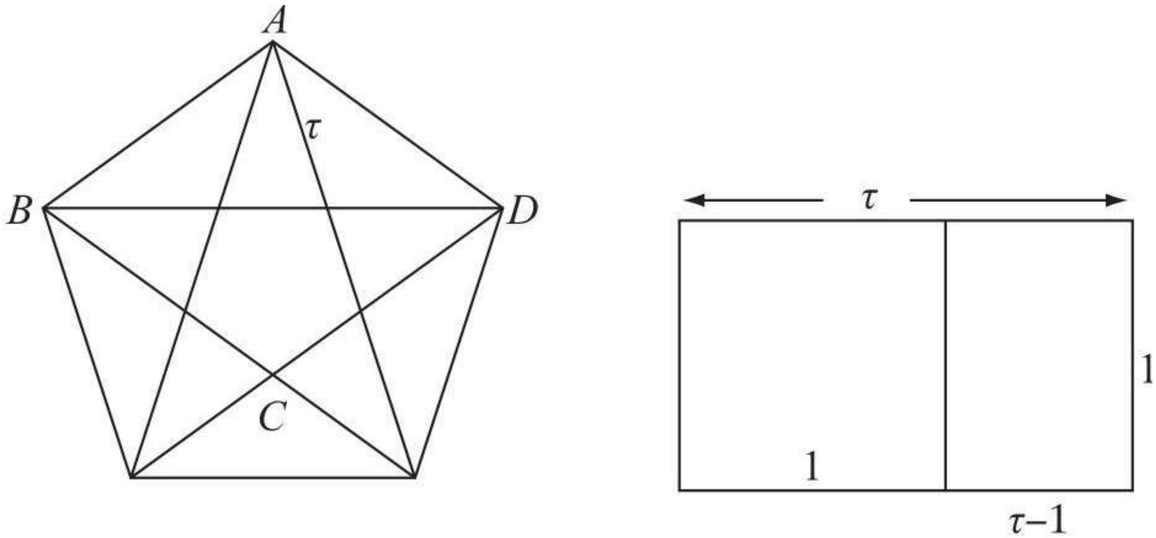


图 4　正五边形和黄金长方形

黄金分割比的一个特征是**自相似性**(selfsimilarity)：一些物体包含自身更小的拷贝。这些物体通常含有比值τ，就像五边形一样。自相似性可以在边长为τ和1的长方形中观察到。这个长方形有一个独特的性质，如果我们从中切下尽可能大的那个正方形(边长为1)，那么留下的那个小长方形正是原长方形的缩小版拷贝。出于这个原因，图4中的图形被称作**黄金长方形**(Golden Rectangle，见图4)。τ的值可以从我们已经给定的性质中推算。我们设原长方形长为τ，宽为1。由于大小两个长方形相似，即它们的长宽之比相等。于是得到等式$\frac{\tau}{1}=\frac{1}{\tau-1}$。将表达式交叉相乘，则得到方程$\tau^2-\tau=1$。用标准方法求解这个**二次方程**(quadratic equation)(含有平方项的方程)，我们得到一个正的根：

$$\tau=\frac{1+\sqrt{5}}{2}=1.618\,033\,9\cdots.$$

另一种获得τ的值的方法是通过它的连分数。连分数能将τ和斐波那契数列直接联系起来，我们将在第7章中探索这一方法。

不断数下去，斐波那契数列看起来就像一个几何数列，它的公比是黄金分割比。正是由于这一性质和它简单的构成规则，斐波那契数列无处不在。

斯特林数和贝尔数

像二项式系数一样，斯特林数[6](Stirling numbers)经常在计数问题中出现。它依赖于两个变量，n和r。斯特林数$S(n, r)$是将一个有n个元素的集合分割成r个子块的方法的个数，每一块都非空，且无关于块的次序和块内部元素的顺序。(严格地说这些称为第二类斯特林数。第一类斯特林数与此相关，但代表了非常不同的东西，即将n个物体排列成r个环的方法总数。)例如，含有元素a，b，c的集合只能以一种方式分成三块：$\{a\}$，$\{b\}$，$\{c\}$；或是以三种方式分成两块，分别是$\{a, b\}$，$\{c\}$和$\{a\}$，$\{b, c\}$和$\{a, c\}$，$\{b\}$；或是以唯一一种方式分成一块：$\{a, b, c\}$。因此$S(3, 1) = 1$，$S(3, 2) = 3$以及$S(3, 3) = 1$。由于将一个n元素集合分割为一块或是n块都只有一种方式，因此$S(n, 1) = 1 = S(n, n)$。如果我们仿照帕斯卡三角形，也将斯特林数放置在一个三角形中，将得到如图5所示的阵列。现在我们来解释这个三角形是如何产生的。

1

1 1

1 3 1

1 7 6 1

1 15 25 10 1

1 31 90 65 15 1

1 63 301 350 140 21 1

⋮

图5　斯特林三角形

这些数也满足了一个递推关系，意味着每个数都联系到阵列中之前的数。就像二项式系数一样，每个斯特林数都可以由它上方的两个数生成，但这次不再是简单的加和。另外，我们看到二项式系数生成的算术三角形具有行对称性，这在斯特林三角形中不再成立。如s(5，2) = 15，然而s(5，4) = 10。不过递推规则依然简单。比如，90这一项等于15+3× 25。这其实揭示了一般情形：要找到三角形中的某个数，取其上方的两个数，**将第二个数乘以当前未知数所在的(自身行内的)列数**，再加上第一个数。(不同于算术三角形，这次列数从1开始计。)用类似的方法，s(5，4) = 10 = 6+4× 1。以上计算规则中只有加粗的部分跟算术三角形的情形不同，但这足以使对斯特林数的研究难度大大超过二项式系数。举个例子，我们推导出了一个简单的、基于阶乘的显式公式，可以计算所有二项式系数。类似地，第n个斐波那契数也有一个基于黄金分割比的幂次的通项公式[7]，但是对于斯特林数，还没有这类公式为人所掌握。

这个递推关系并不难解释。我们的推理类似于二项式系数的递归，给出的递推式也与那里的形式一致，只是相差了一个乘数。为了将一个n元素集合分成r个非空子块，我们可以采取两种不同的方法。

我们可以取集合的前n-1个元素，并将其分成r-1个非空块，共有S(n-1，r-1)种方法，最后一个元素将构成第r个块。或者，我们可以将集合的前n-1个元素分成r个非空块，这有S(n-1，r)种方法，接下来再决定将最后一个元素放进r个块的哪一个，这就需要用r乘以目前的方法数。因此有S(n，r) =S(n-1，r-1)+rS(n-1，r)，这里n= 2，3，… .

用这个递归公式，我们就可以基于上一行计算每一行斯特林三角形的值。如，设n = 7和r = 5我们得到：

S(7， 5) = S(6， 4)+5S(6， 5) = 65+5× 15 = 65+75 = 140 .

可以用定义直接计算S(n， 2)和S(n， n-1)。将一个n元素集合分割成两个子集，这一过程可以由一个长度为n的二进制字符串 来描述，其中1表示相应位置的元素在第一个子集中，而0则表示该元素属于第二个子集(类似于我们证明n元素集合的子集个数是2n)。于是，有2n个这样的有序子集对。但是因为分割与块的次序无关，我们将这个数目除以2，便可得到将n元素集合分成两个子集的方法个数，即2^{n-1}。最后，还需要从中减掉1，去除掉其中一个子集为空的情况。因此，S(n， 2) = 2^{n-1}-1。对照图5，你可以检查看看，这个公式给出了右上到左下方向第二对角线上的数，即1， 3， 7， 15，31， 63，…。

另一种极限情况是将一个n元素集合分成n-1块，这等价于选唯一一个含有两个元素的子块。选择的方法个数是 $C(n, 2)=\frac{1}{2}n(n-1)$ ——第n-1个三角形数(见图5从左上往右下的第二个对角线)。

算术三角形中任意一行的和给出了对应2的幂次——一个集合的子集数量。类似地，斯特林三角形的第n行的和给出的是将n个物体分成任意个数子块的方法总数，这被称作第n个**贝尔数**(Bell number)。**分拆数**

另一方面，如果待分割集合中的n个物体是一模一样、无法区分开的，将整个集合分拆为子块的方法数就变得小得多了。这称为第n个**分拆数**[8](partition number)。每一个特定的分拆对应于将n写成一些不考虑次序的正整数的和。例如，1+1+1+1+1是5的一个分拆，还有6种其他分拆，因为我们还可以将5表示成1+1+1+2， 1+2+2，1+1+3，2+3， 1+4，或直接就是5。因此第5个分拆数为7(对比第5个贝尔数，后者可由斯特林三角形计算，即1+15+25+10+1 = 52)。没有简单的精确公式可以计算第n个分拆数，但有一个复杂的公式。该公式基于印度天才数学家斯里尼瓦瑟・拉马努金(Srinivasa Ramanujan)给出的一个优美近似。

分拆具有一种简单的对称性，那就是将n分拆为m个数的方案数等于将n分拆后所得数中最大值恰为m的方法个数。要想看到这一结论的正确性，一种方法是通过分拆的**费勒斯图像**(Ferrar’s graph)——或称**杨表**[9](Young diagram)。这个图表并不神秘，无非是将分拆表示成元素逐行减少的点阵。

在图6的例子中，我们把17拆分为5+4+4+2+1+1。注意到从左向右各列也以降序排布。如果我们绕着左上到右下的对角线翻转整个点阵，我们就得到图示的第二个费勒斯图像，这个图像可以阐释为分拆17 = 6+4+3+3+1。对第二个图像做同样的翻转，又会回到第一个图像。我们称这两种分拆互为**共轭**(dual)。这一对称性表明两种类型的分拆方案数是相等的：一种是m为结果中最大的数的分拆(即顶端行有m个点)，它的共轭是一个有m行的分拆，即拆分为m个数。例如，将17拆分为6个数的方案数，等于将17拆分使得最大数为6的方案数。

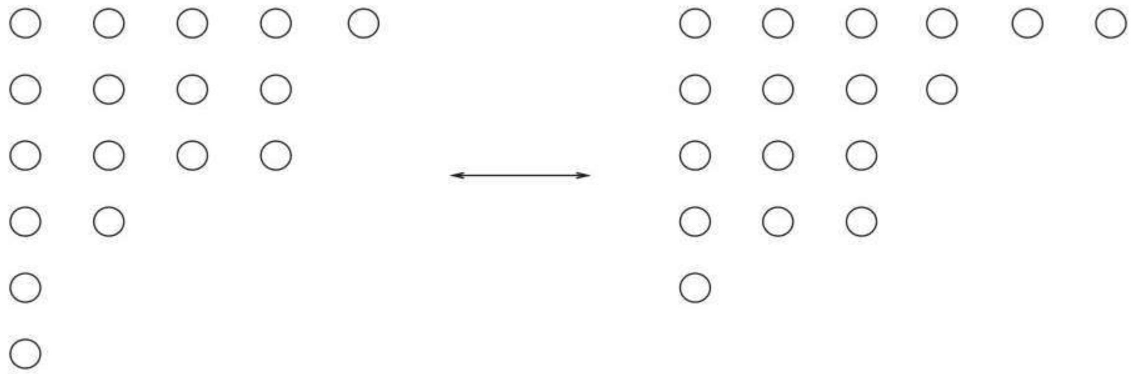


图6　共轭分拆，17＝5+4+4+2+1+1＝6+4+3+3+1

数学家们总是会寻找经常在计数问题中出现的这类对称性。这方面的另一个例子是伯特兰-惠特沃斯选票问题(Bertrand-Whitworth Ballot Problem)。假设有两位候选人，经过唱票，获胜一方得到p张选票，而失败一方有q张票。一个巧妙的几何方法运用所谓的**反射原理**(Reflection Principle)，证明了以下结论：唱票过程给出一种候选人得票情况的排列，从头至尾胜者得票数都领先于败者的排列的个数占所有可能得票排列总数的比例，等于胜者的净胜票数除以总票数$\frac{p-q}{p+q}$。这恰巧还等于直到最后一票唱完，胜者净胜票数才增加到$p-q$的可能排列个数占所有可能得票排列总数的比例。两个比例之所以相等，是因为这两种类型的得票排列恰好互为共轭。在一种类型排列的唱票中，将得票次序颠倒过来，即可得到第二种类型排列的唱票，反之亦然。

冰雹数

尽管不是一种计数工具，冰雹数(hailstone numbers)仍然耐人寻味，因为它也是被递推定义的，不过这个数列更像我们在第3章中遇到的真因数和数列。以下问题有好几个名字，**考拉兹算法**(Collatz algorithm)、**叙拉古问题**(Syracuse problem)，或者有时就叫**$3n+1$问**

题[10]。它基于一个简单的观察，即从任意数n开始，经过以下步骤似乎最终总是得到1：若n是偶数，将其除以2；若n是奇数，用3n+1代替它。例如，从n = 7开始，我们按照这些规则得到以下数列：

$$7\to 22\to 11\to 34\to 17\to 52\to 26\to 13\to 40$$

$$\to 20\to 10\to 5\to 16\to 8\to 4\to 2\to 1.$$

于是这一猜想对n = 7成立。实际上该猜想已经被证实对于直到一万亿以上的某个数都是正确的[11]。但是如果你乱动了规则，事情就大不一样了。比如，用3n-1代替3n+1会导致一个循环：

$$7\to 20\to 10\to 5\to 14\to 7\to \cdots.$$

考拉兹算法产生的数列类似于冰雹，在很长一段时间内它们的取值大起大落，但最终似乎总会降落到地面。在前1000个正整数中，有超过350个数上升到最大高度9232，而后则迅速跌落到1。一旦你遇到2的幂次，就会最终得到1，因为之后这些数不会经历任何爬升，只能一路衰减到底。

所有这些奇妙的特性都可以在基于冰雹数列的图像中被发现，这让人不禁联想起数学和物理中出现的其他混沌现象。在搜索引擎中输入“冰雹数”，你会找到一大堆信息，这些信息往往看上去奇妙异常，有时候是猜想性质的，但通常不能给出一个确定的结论。

[1] 即等比数列。

[2] 该公式中$r\neq 1$。

[3] 即杨辉三角形。

[4] 朱世杰，字汉卿，号松庭，元代数学家和教育家。

[5] 尤金·查理·卡特兰(Eugène Charles Catalan)，法国和比利时数学家。

[6] 詹姆斯·斯特林(James Stirling)，苏格兰数学家。埃里克·坦普尔·贝尔(Eric Temple Bell)，英国数学家、科幻小说家。

[7] 如果可以将数列{a_n}的第n项用一个含参数n的式子表示出来，则称该式为该数列的通项公式。

[8] 又称拆分数或分割数。

[9] 又称杨氏矩阵。

[10] 又名考拉兹猜想、叙拉古猜想、3n+1猜想、奇偶归一猜想、冰雹猜想、角谷猜想、哈塞猜想、乌拉姆猜想。

[11] 世界各地的数学爱好者们利用分布式计算，不断地提高这个极限。已知的猜想成立的最大整数似乎已超过10^{20}(一亿亿的一百倍)，而仍未找到反例。有兴趣的读者可以参考以下网站：https://boinc.thesonntags.com/collatz/，http://www.ericr.nl/wondrous/。

06 数之冰山的水下部分 Below the Waterl ine of the Number Iceberg

引言

用于数数的自然数1， 2， 3，…只是全部数的冰山一角，当然，这也是我们首先发现的部分。一时间我们或许会相信，冰山除了这个角就没有什么其他东西了，尤其是如果我们不去看水面以下部分的话。在这一章的旅程中，我们首先引入负整数。再与正和负的分数(fraction)结合在一起，我们就得到了名为**有理数**(rational number)的集合。这个数集常常被看作沿着**数轴**(number line)分布，其中正数位于0的右边，而与它们对应的负数则在左边形成镜像。然而，人们发现数轴上还有其他不能表示为分数的数，其中两个例子是$\sqrt{2}$和π 。**实数**(real number)是数轴上所有数的统称，它们是所有能被小数展开式所表示的数，如图7。

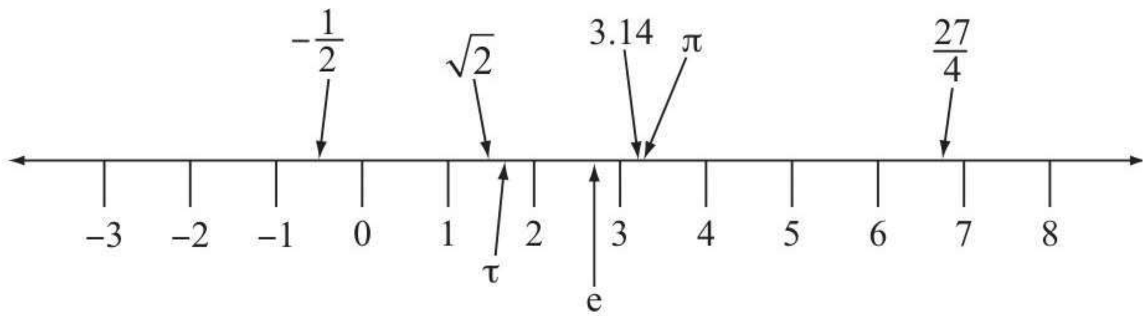


图7　数轴上在0附近的中央部分

不过，人们在19世纪最伟大的成就之一是充分意识到数域其实不是一维的，而是二维的。复数(complex number)构成的平面才是大部分数学论辩的天然场地。这个结论是数学家和科学家通过解决问题才

意识到的——为了能够开展研究，解决现实中的问题，有必要扩展数的边界。虽然很多问题似乎都只跟普通的自然数有关。关于这个额外的维度是怎样出现的，我们将在本章的末尾做出解释，并在第8章中进一步探讨这个话题。

加和减

整数指代所有“整的”数组成的集合，包括正的、负的以及0。这个集合通常用字母Z[1]来表示，它向两边无穷延伸：

{…，-4，-3，-2，-1， 0， 1， 2， 3， 4，…}

我们常把整数看作水平数轴上等距的点，它们按以上的次序排列。为了能用整数运算，下面总结了我们需要知道的额外的规则：

(a)加上或减去一个负整数，-m：加的时候我们向左移动m个位置，减的时候我们向右移动m个位置。

(b)乘以整数-m：我们将原整数乘以m，接着再改变符号。

换句话说，加上或者减去负数的方向和正数情况下的相反，而将一个数乘以-1则会使得它的符号反转。比如：8+(-11) = -3，3×(-8) = -24，(-1)×(-1) = 1。

你无需为最后那个式子困扰。首先，一个负数乘以一个正数得到负数，这是合理的。因为当债务(负的量)产生了利息(一个大于1的正乘数)，结果会是更重的债务，也就是说一个值更大的负数。这一点我们都很清楚。一个负数乘以一个负数，应该给出相反的结果，即一个正数，这样才与前面的一致。我们甚至可以给负负得正这个事实一个严格的证明。它基于这样的假设：我们希望扩展的整数系统包含了原来的自然数，并且继续遵守所有代数运算的普通规则。事实上，两个

负数的积可以从任何数乘以0等于0推出。(这个结论也不是一个假设，而是代数法则的必然结果。)我们现在有：

$$-1\times(-1+1)=-1\times 0=0\ .$$

将括号拆开，我们看到要想左边等于0，(-1)×(-1)必须与(-1)×1反号，换句话说(-1)×(-1) = 1。

分数和有理数

就像减法导致了负数的出现，除法运算同样将我们带到自然数集之外，进入了分数的领域。然而，我们遇到的新的算术具有不同的性质。当相加或相减时，分母(denominator)(位于下方的数)不同的分数是不相容的。参与运算的分数需要被表示为分母相同的形式，才能求和。乘法相对来说简单一些，因为我们只需要将分子(numerator)(位于上方的数)和分母分别乘在一起，就能得到答案。除法是乘法的逆运算，因此除以n相当于乘以它的倒数(reciprocal)$\frac{1}{n}$。这也适用于一般的分数，要除以分数$\frac{m}{n}$，我们乘上它的倒数$\frac{n}{m}$，因为这个步骤反转了乘以$\frac{m}{n}$的效果。

古埃及人偏爱**单位分数**(unit fraction)，即那些整数的倒数：$\frac{1}{2},\frac{1}{3},\frac{1}{4}$等。(虽然他们为$\frac{2}{3}$保留了一个特殊符号)而像$\frac{3}{4}$这样的分数不被认为是有意义的独立存在的实体。他们将这个量记作两个倒数之和：$\frac{3}{4}=\frac{1}{2}+\frac{1}{4}$。(这里使用的分数的记号当然是现代欧洲式的，其源头是希腊数学)古埃及人坚信任何分数都可以写成几个**不同的**单位分数之和。这个结论并不显然，不过这确实总是可以做到的。解释这个结论的过程可以帮助你复习一下处理分数的技巧。

假设有一个分数，比如$\frac{9}{20}$，如果你想找到它的一个埃及式分解，只需要从中减去能减去的最大的单位分数，重复这一过程直至剩下的数自己也是一个单位分数。这总是可行的，并且涉及的分数的数目不可能超过原来分数的分子。这是因为在每一步，剩下的分数的分子总是比之前一个数的分子小——并不显然，但确实是真的。具体到这个例子里，第一步：

$$\frac{9}{20}-\frac{1}{3}=\frac{27}{60}-\frac{20}{60}=\frac{7}{60}.$$

接着，我们发现小于$\frac{7}{60}$的最大单位分数是$\frac{1}{9}$。（要想检验的话，比较交叉相乘的结果：因为1× 60=60＜63=7× 9，所以$\frac{1}{9}<\frac{7}{60}$。）再次相减，我们看到

$$\frac{9}{20}-\frac{1}{3}-\frac{1}{9}=\frac{7}{60}-\frac{1}{9}=\frac{21}{180}-\frac{20}{180}=\frac{1}{180}.$$

因此，我们找到了原分数的埃及分解：

$$\frac{9}{20}=\frac{1}{3}+\frac{1}{9}+\frac{1}{180}.$$

虽然这个减去最大单位分数的贪心法总是可行的，但它不一定给出最短的分解，因为即便是这个例子里，我们也能看出$\frac{9}{20}=\frac{1}{4}+\frac{1}{5}$。不过，$\frac{9}{20}$的这个双分数分解可以用艾赫米姆莎草纸（Akhmim papyrus）上所记载的方法找到。艾赫米姆莎草纸是在尼罗河（Nile）上的城市艾赫米姆找

到的一张希腊羊皮纸，它的年代可以追溯到公元500—公元800年。使用现代记号，这个技巧可以表达为以下的代数恒等式，这个等式很好验证：

$$\frac{m}{pq}=\frac{m}{p(p+q)}+\frac{m}{q(p+q)}.$$

将这个等式应用于$m=9$，$p=4$，$q=5$，我们立即得到

$$\frac{9}{20}=\frac{9}{4\times 9}+\frac{9}{5\times 9}=\frac{1}{4}+\frac{1}{5}.$$

哪怕是简单的分数加法，你也可能多年没有做过了，因为现在生活中几乎所有运算都是通过十进制格式的小数来完成的。十进制分数在古代中国和中世纪阿拉伯国家就已出现，但直到16世纪下半叶才在欧洲广泛传播，那时人们花了很大精力来改进计算的实用性。然而，为了使用这种十进制形式，我们也付出了一定的代价。在通常的以十为底的算术中，我们利用了任何数都可以写成十的各次幂的倍数之和这一事实。当用十进制表示分数的时候，我们试图将一个数写成$\frac{1}{10}=0.1$的各次幂之和。不幸的是，即便是对于$\frac{1}{3}$这样非常简单的分数，这也没法做到，它的小数展开会无休止地进行下去：$\frac{1}{3}=0.333\cdots$。我们得尊重这一事实，实际使用时需要在一定的数位之后(取决于所需要的精确度)截断小数展开式。这样我们就用**有限小数**(terminating decimal)来近似精确的分数，以此方式做出妥协。相比于在标准的十进制参照系下开展数的工作所带来的方便，这点不精确显得微不足道。不妨认为，十进制小数展开是最接近于使所有分数都具有一个共同分母的方法。

不过自然有人会问，哪些分数会只有有限位的展开(而哪些没有)？答案是，不太多。通常，一个分数的小数展开会进入一个循环，比方说$\frac{3}{22}=0.136\ 363\ 6\cdots$，36这部分会永远重复下去。每个分数都会以某种方式产生循环。对于有限小数，比如$\frac{1}{2}=0.5$，重复出现的部分仅仅是一串无穷的0；$\frac{1}{2}=0.500\ 0\cdots$，于是我们不再提及。在任何情况下，循环部分的长度都不会大于分母的值减一。这可以通过考虑完整的除法算式来看到：假如分母是n，那么除法中每一步后的余数是0， 1，…， $n-1$中的一个。如果在某一步余数为0，则除法结束，小数展开也就终止了。例如，$\frac{11}{40}$严格等于0.275。否则的话，除法将永远进行下去。某个余数将不可避免地重复，一旦重复出现，我们就会被强制拉进一个循环。因此就出现了一个循环模式，其周期长度不超过$n-1$。当且仅当分母是若干2和5的乘积时，小数展开是有限位的，而2和5正是我们的底数10的素因数。但若是分母包含任何其他因数，小数便会是无穷的。例如，分母为16， 40和50的分数都是有限小数，但像$\frac{1}{14}$和$\frac{1}{15}$这样的分数则是无穷小数，因为它们的分母分别含有因数7和3，这迫使小数展开进入无穷循环。

但是，以上这些恰恰显示了，一个分数的小数展开是否有限并非取决于这个数本身，而是依赖于它和你所取的底数的关系。比方说，假如我们使用**三进制**(ternary)(底数为3)，那么0.1就表示$\frac{1}{3}$，因为小数点后的1代表$\frac{1}{3}$，而不是像在十进制展开里那样代表$\frac{1}{10}$。

将一个无穷循环小数反过来转换成一个分数也十分简单，这表明在分数和循环小数之间存在一一对应关系，而我们可以选择一种最适合当前用途的表示方法。下面我们举一个简单的例子：设a=0.212 121…。因为循环部分的周期为2，因此我们可以通过乘上10^2 = 100将其简化。正如你看到的，我们得到100a=21.212 121…。这样做是为了让a和100a的循环小数部分相等，当两个数相减时，这部分可以互相消去。这使得我们能推断出99a=21，于是$a=\frac{21}{99}=\frac{7}{33}$

这样的技巧常常被用于简化包含无穷重复过程的表达式。比如，考虑下面这个令人生畏的式子：

$$a=\sqrt{2\sqrt{5\sqrt{2\sqrt{5\cdots}}}}\ .$$

通过求平方，接着再一次平方，左侧变成a^4，而右侧的表达式变成：

$$a^4=2^2\times 5\times\sqrt{2\sqrt{5\sqrt{2\sqrt{5\cdots}}}}\ .$$

由于5后面跟着的正是a的表达式，我们推知$a^4=20a$，于是$a^3=20$，或者a是20的立方根——如果你更喜欢这样说。在第7章中我们会再次用到这个技巧，那里我们将介绍所谓的连分数(continued fraction)。

分数这一类别是否提供了我们可能需要的所有数了呢？正如之前提到的，所有分数以及它们的负数的总合，形成了被称为有理数的集合，也就是由整的数和它们之间的比值所产生的所有的数。它们对于算术来说是足够的，这意味着，涉及加、减、乘、除四种基本运算的任何结果都不会将你带出有理数的范围。如果我们对此感到满意，那么这个数集就是我们所需的。不过，在下面的小节，我们来解释为什么像上面的a那样的数不是有理的。

无理数

当特指一个数a的时候，**无理的**(irrational)这个词仅仅表示这个数不是有理数，也就是说它**不能**被写成一个分数。无理数的首次发现是在很久以前的古希腊。毕达哥拉斯理解了$\sqrt{2}$的无理特性。古希腊人并没有从小数展开的角度来思考问题，他们欣然接受只要是用直尺和圆规作出来的几何图形，其中的长度就代表了真实的量。特别地，勾股定理(Pythagoras' Theorem)告诉我们，一个两条短边长度都为1的直角三角形，其长边长恰好等于$\sqrt{2}$。

毕达哥拉斯能够证明，$\sqrt{2}$不等于任何分数，这显示了无理数的确是存在的。特别是你测定一个正方形的边长后，就没法再用同样的单位精确地测定它的对角线的长度。不然对角线长度就会是边长的分数倍，即$\sqrt{2}$等于这个分数。这两个长度从根本上来说是不相容的，或者用古典文献的话来描述，就是**不可公度的**[2](incommensurable)。π也是如此，它大约等于分数$\frac{22}{7}$，但又不完全一样，并且与你能提出的任何候选分数都不同。(不过“两个1，两个3，两个5”这个好记的比例$\frac{355}{113}=3.1415929\cdots$相当精确地近似了$\pi$的值，直到百万分位。)虽说证明$\pi$是无理数非常困难，$\sqrt{2}$的问题却可以用一个简单的反证法轻松解决。首先，我们注意到对于任意数c，c^2的因数中2的最高次幂的指数等于c的因数中2的最高次幂的指数的两倍。于是任何平方数的因数中2的最高次幂的指数必须是一个偶数。例如，$24 = 2^3 \times 3$，而$576 = 24^2 = 2^6 \times 3^2$。这里，当我们取24的平方的时候，因数中2的最高次幂的指数的确从3变成了6。这个结论总是成立的，并且不仅仅适用于2的幂，也适用于原数的任意素因数的幂。

现在假设$\sqrt{2}$等于分数$\frac{a}{b}$。将这个方程的两边平方，我们推出$2=\frac{a^2}{b^2}$，于是得到$2b^2 = a^2$。根据之前的假设，方程右侧的因数中2的最高次幂的指数为偶，而左侧的因数中2的最高次幂的指数为奇(由于存在那个额外的2)。这表明该方程是没有意义的，因此最开始将$\sqrt{2}$写成一个分数

一定是不可能的。就像毕达哥拉斯一样，我们已经和无理数面对面了。

使用类似的推理，我们能够证明，一般取一个数的平方根(或是立方根甚至是更高次方根)的时候，答案如果不是一个整数，就总是一个无理数。这就解释了当你计算方根的时候，为什么你的计算器上显示的小数从来都没有循环的迹象。

毕达哥拉斯发现要想发展他的数学，他需要一个比分数更大的数域。古希腊人认为，如果可以从一个标准单位长区间出发，仅用一把直尺(并非带刻度的尺子，仅仅是一根直棒)和一只圆规作出一根线段，那么这根线段的长度所代表的数就是“真实的”。人们发现，虽然平方根运算的确引入了无理数，但是数集本身并没有从有理数出去太远。我们称之为欧几里得数(Euclidean number)的集合，包括所有那些可以从1出发通过四则运算，以及任意次平方根得到的数。比如，数$\sqrt{7-\sqrt{\frac{4}{3}}}$就是一个这种类型的数。但甚至是立方根也超过了欧几里得工具的范围。这构成了或许是数学里第一个伟大的未解决问题的基础。三个所谓提洛问题(Delian problems)中的第一个，就是用尺规构造2的立方根。传说中，提洛岛(Delos)的居民找到德尔菲的祭司(Oracle of Delphi)，询问应该如何做才能从雅典(Athens)驱除瘟疫，而神谕给他们的正是这一任务——问题的表述是将一个正方体祭坛的体积恰好增大一倍。

这个问题在古典时代(classical times)一直无人问津。直到1837年皮埃尔·汪策尔(Pierre Wantzel)才将其“盖棺定论”：2的立方根在欧几里得工具所能到达的范围之外。这么晚是因为我们需要一种精确的代数来描述古典工具能达到的极限，这样才能看出2的立方根从根本上讲是一种不同类型的数。实际上，最后这可以归结为证明用平方

根和有理数永远不可能造出立方根。这样说的话，这个不可能性听起来似乎更合理一些了。当然，这还不能构成一条证明。

超越数

无理数中还存在着神秘的超越数(transcendental)家族。这些数**不能**由普通的算术运算或是求方根得出。在给出精确的定义之前，我们先介绍与之互补的集合，**代数数**(algebraic number)，其中每个数都是一个拥有整数系数的多项式方程的解。例如，$x^5-3x+1 = 0$就是这样的一个方程。**超越数**则被定义为非代数数。

到底有没有这种数呢？答案完全不明显。不过，它们确实是存在的，只是它们的社群十分隐秘，其中每个成员都对自己的会员身份讳莫如深。比如π 这个数就是超越数的一例，但这不是一目了然的事情。在下一章中，我们将要探索无限集合的性质，那时候我们会解释为什么“大部分”数都是超越的，我们会严密地阐释这个“大部分”的含义。

就目前而言，我介绍一下大概是所有超越数中最著名的一个，数e=2.718 28…。这个数在高等数学和微积分中经常出现，它是所谓的自然对数(natural logarithm)的底。自然对数告诉你反比例函数图像下方的面积。当你将两个连续整数的比值，$\frac{n+1}{n}(=1+\frac{1}{n})$取$n$次幂，就会得到一列依赖于$n$的数，这个数列的极限值也是$e$。用你的计算器求$(\frac{129}{128})^{128}$——你也可以对它应用“快速求幂”，计算$\frac{129}{128}$，再平方7次，因为$2^7$=128。

在复利问题中，逐步缩短付款时段，从每年改成每月，再改成每日，等等。我们考虑复利的利率会趋向于什么样的极限值，这时候上面的数列就会出现。为了更好地说明这一点，假设利息以100%的年利率计算，一年内分n次付款，并计算复利。这意味着你的初始投资会乘以$(1+\frac{1}{n})$，在一年的时间内总共被乘n次。于是你的本金便会被乘以因数$(1+\frac{1}{n})^n$

。利息支付得越频繁，你赚得就越多，因为随着n越来越大，你对已获得的利息也越来越早开始收取利息。不过，随着n的增大，其等效的年利率(annual percentage rate， APR)并不会无限地增长，而是会趋近于一个值——数学家们称之为上限(upper limit)。在连续复利的情形下，这个极限会乘在你的本金上，它是随着n增大时下面这个数的极限值

$$\left(1+\frac{1}{n}\right)^n \to e = 2.718\ 28\cdots.$$

另一种产生神秘的e的方法是将阶乘的倒数相加。这也是一种以很高精度计算e的途径，因为这个级数(series[3])的各项迅速趋近于0，于是级数本身会很快收敛：

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots.$$

有了这个表达式，我们就能用一种较简单的反证法证明e是一个无理数，这里我们概括地介绍一下。假设前述e的级数等于一个分数$\frac{p}{q}$，然后我们在两边同乘以q!。于是左侧变成了一个整数，但是右边则先是一些整数项，后面跟着一个由非整数项组成的无穷数列。通过与一个简单的几何级数做比较，我们可以推出这条“尾巴”加起来小于1，于是右侧不可能是一个整数，这样我们就得到了反证需要的矛盾。如果要证明e不仅仅是无理数，还是超越数，那就需要工作量就大得多的了。

e和阶乘的关系也在一个漂亮的公式中体现出来，这个公式是由苏格兰数学家詹姆斯·斯特林(James Stirling)找到的——斯特林数(见第5章)就是以他的名字命名的。他证明了随着n的增大，n!的值可以由表达式$\sqrt{2\pi n}(\frac{n}{e})^n$越来越精确地逼近。

由于e会以各种不同又很简单的方式出现，所以它渗透了数学的各个分支，并且常常是那些你意料之外的地方。例如，取两幅彻底洗开的扑克牌，翻开每副牌最上面的那张进行比较。重复这个过程，直到用完两副牌。那么，在某一步完美契合的机会有多大呢？也就是说，在某一次翻牌的时候，摸出的两张牌一模一样，无论是梅花7、红桃Q，还是任何其他牌。最后人们发现，做这样的实验，至少出现一次完美契合的实验次数占总次数的比例几乎与$\frac{1}{e}$相等，也就是大约36.8%。我们可以应用所谓的**容斥原理**(inclusion-exclusion principle)来推出这个结果，这是通过将一系列代表修正和反向修正的项相加得来的。在这个例子中，这条原理产生了阶乘倒数的级数，只不过这一次这些项的符号交替变化，这个级数收敛到$\frac{1}{e}$。

实的和虚的

本书的前五章主要都在和正整数打交道。我们强调了整数的因数分解性质，这引导我们去考虑不具有真正分解的数——也就是素数，这个集合在现代密码学中占据了举足轻重的位置。我们还了解了一些具体类型的数，比如和完美数有紧密联系的梅森素数。我们耐心地介绍了一些特殊的整数，对某些自然出现的集合计数的时候，它们扮演重要的角色。在所有这些数当中，大背景都是整数系统，即自然数、它们的负值以及0。

在这一章我们走出了整数的范围，首先是进入有理数的地界(分数，包括正的和负的)，接着又走进了无理数。在无理数这个类别中我们认识了超越数。所有这一切背后的基础是**实数**系统，实数可以看作所有可能的小数展开式的集合。任何正实数都可以用$r= n.a_1a_2\cdots$的形式来表示，这里n是一个非负整数，小数点后面跟着一串 无穷多的数字组成的尾巴。如果这条尾巴最终进入一个循环，那么r其实是有理数，而我们已经介绍过怎样将这个表达式转换成一个普通的分数。如果没有进入循环，那么r是无理的。因此，实数包含了这两种不同的类型，有理数和无理数。

在我们的数学想象中，我们常常将实数看作对应于数轴上所有的点，从0向外看去，右边是正数，而左边是负数。这给了我们一个对称的图像，负实数构成正实数的镜像。这一对称性在加法和减法运算中得以保留，但不适用于乘法。一旦我们进入乘法的范畴，正负数便不再拥有同等的地位，因为数1被赋予了其他数都没有的性质，它是**乘法单位量**(multiplicative identity)。意思是说对于任意实数r，都有$1\times r=r\times 1=r$。乘以1使得任何数原地不动，但是相反地，乘以-1会将一个数和它在0的另一边的镜像对调。一旦乘法进入我们的视野，正数和负数在性质上的本质不同便显现了出来。特别是，负数在实数系统中不具有平方根，因为任何实数的平方总是大于或等于0。

这一情形恰恰呼唤着我们的虚数(imaginary number)的登场。在最后一章中我们会重拾这一话题。现在，让我们只做一些介绍性的点评。

这些数的出现是因为人们要寻找简单的多项式方程的解。特别是，由于任意实数的平方从不为负，我们找不到方程$x^2=-1$的解。数学家们并没有被这吓住，而是发明了一个解，记为i，并且赋予了它这样的性质：$i^2 = -1$。这乍看起来似乎很随便，但它其实跟我们之前习以

为常的做法并没有太多不同。原因在于，自然数固然很重要，但为了顺利处理有关数的一般问题，我们还是得引入更为广泛的有理数系统，即正负分数以及0的集合。但是接下来，我们又发现自己没有方程$x^2=2$的解，因为我们已经证明，一个有理数的平方不可能恰好为2。为了应付这一状况，我们只能“发明”$\sqrt{2}$。在这里，我们当然也可以采取另外一种态度，声称我们已证明2的平方根就是不存在而已，故事便在此终结。然而，很少有人会止步于此。古希腊人尤其不会喜欢这样一个烂摊子。因为他们能够用尺规构造出代表$\sqrt{2}$的长度，所以在他们的思维方式里，这个数是真实存在的，任何否认这一点的数学系统是不完善的。

我们也可能出于另一个完全不同的原因而同意毕达哥拉斯的看法。我们或许会说，通过小数展开$\sqrt{2}$=1.414 213…，我们能以无限精确地逼近$\sqrt{2}$，因此$\sqrt{2}$正是被这个完整的无穷展开所代表的数。现代人可能更会被这个说法所打动，于是他由此出发坚持认为数的系统需要扩展到有理数以外。

但第一眼看起来，到了$\sqrt{-1}$这里，事情或许又有所不同，因为在我们习惯性地称为“真实的”数集中，似乎并不用担心这个数的缺席。但是随着我们的数学推进了一小点，对虚数的需求就变得十分急迫了。当我们对事物数学属性的理解增长到一定程度，一开始对虚数的不情愿都烟消云散了。

在16世纪，意大利数学家们通过推广二次方程的解法，学会了如何解三次和四次多项式方程，此时这个问题第一次全面浮现出来。后来所说的卡尔达诺法[4](Cardano method)中，虽然到最后方程的解是正整数，但计算过程常常涉及负数的平方根。从那时起人们逐步发现，使用**复数**可以让许多数学计算得以开展。复数是形如$a+bi$的数，这里a和b都是普通的实数。例如，在18世纪，欧拉发现并应用了$e^{i\pi}$ =

-1这个小小的出人意料的等式，每一个第一次见到它的人都会禁不住惊讶。

在19世纪早期，韦塞尔[5](Wessel)和阿尔冈[6](Argand)研究了复数的几何解释——即坐标平面(标准的xy-坐标)上的点，在这之后“虚的”这一术语被普遍接受为数学用语。将复数x+iy和具有坐标(x，y)的点对应起来，这使得我们能够通过平面上的点的行为来研究复数的行为，事实证明这是极富启发性的。关于所谓的**复变量**(complex variable)的理论，研究的是依赖于复数——而不仅仅是实数——的函数。经由奥古斯丁·柯西(Augustin Cauchy)的发展，这一理论枝繁叶茂。它现在已经成为数学的一块基石，并且为电信号理论提供了数学基础，而整个X射线衍射领域完全建立在复数的基础上。人们已经证明这些数拥有实实在在的意义，除此之外，这个系统还是完备的，因为**每一个**多项式在复数系统中都有一组完整的解。我们会在最终章中回到这些话题。不过，在那之前，让我们在下一章中先更近距离地观察一下实数轴的无穷特性。

[1] 也经常用符号$\mathbb{Z}$表示，后文的C也常写成$\mathbb{C}$。

[2] 或称不可通约的。

[3] 此处的series作为“级数”的意思，是无穷数列各项之和。注意这与之前series的意思不同，那里series指数列。

[4] 吉罗拉莫·卡尔达诺(Girolamo Cardano)，意大利学者。卡尔达诺法是求解一般三次方程的方法。

[5] 卡斯帕尔·韦塞尔(Caspar Wessel)，挪威、丹麦数学家。

[6] 让-罗贝尔·阿尔冈(Jean-Robert Argand)，法国业余数学家。

07 向无穷和更远出发！To Infinity and Beyond!

无穷之中的无穷

伽利略·伽利雷(Galileo Galilei)是16世纪意大利伟大的博学家，他第一个提醒我们注意，无限集合和有限集合的性质有着根本上的不同。正如本书第一页所暗示的，如果一个**有限**集合的元素可以和另一个集合的部分元素配对，那么第一个集合的大小就小于第二个。然而，与此相反，我们可以用这种方法将无限集合与自身的**子集**(subset)对应起来(这里子集这个术语表示原集合内包含的一个集合)。要理解这一点，我们都不需要超出自然数数列1， 2， 3， 4，…。我们可以轻易地举出这个集合的任意多个子集，它们都形成无限集合，从而与全集形成一一对应的关系(见图8)：奇数1， 3， 5，7，…，平方数1， 4， 9， 16，…，以及不那么明显的素数2， 3， 5，7，…。并且在每种情况下，其相应的补集(complementary set)，即偶数、非平方数及合数，也都是无穷的。

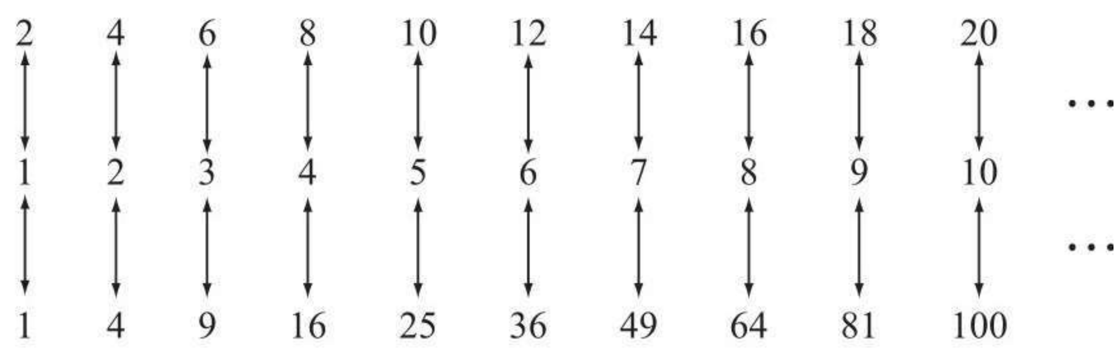


图8 与自然数配对的偶数和平方数

希尔伯特旅馆(Hilbert Hotel)

大卫·希尔伯特(David Hilbert)是他那个年代最杰出的德国数学家。这家十分奇特的旅馆总是与他的名字联系在一起，它生动地体现了无限集合的奇妙性质。这家旅馆的主要特点是它有无穷多房间，用1， 2， 3，…来编号，还夸下海口，希尔伯特旅馆总是有空房。

一天晚上，这家旅馆却客满了，也就是说每一间房里都住了一位客人。令前台服务员沮丧的是，又一个客人出现了，并要住一间房。幸好经理出手了，他把服务员拉到一边，教他如何处理这种情况，这才避免了尴尬的局面。“让1号房的客人搬到2号房，”他说，“2号房的搬到3号，以此类推。换句话说，我们通知整个旅馆，要求n号房的客人换到n+1号房中，这样1号房就可以空出来给这位客人了!”

因此你看，在希尔伯特旅馆的确总是**有**空间的。但是到底有多大空间呢？

第二天晚上，服务员面临着一个类似但更加考验人的情形。这一次，一艘载着1729名乘客的太空飞船抵达了，所有人都闹着要在已经客满的旅馆中各自住进一间房。不过前台服务员从前一晚吸取了经验，知道怎样推广之前的方法来应付这个额外的旅行团。他让1号房的人去1730号房，2号房的换到1731号房，如此这般。他发出了通知，要求n号房的客人搬进n+1729号房间。这就让1～1729号房都空了出来，给新来的人入住。我们的服务员为自己独立处理了昨晚问题的新版本感到骄傲——他这么想也是完全合理的。

然而，最后一天晚上，服务员再次面临同样的情况——一家客满的旅馆，但是这次让他惊恐的是，不是仅仅出现了几名额外的客人，而是一辆无限大的星际快车，载着无穷多的旅客，每个自然数都对应一位客人。焦头烂额的宾馆服务员告诉司机，这家旅馆已经满员，看

不出有什么办法能容纳车上的全部旅客。他或许可以塞下一两个，甚至也可以是任意有限个数，但绝无可能是无穷多的额外客人。这就是不可能!

一场无限大的骚乱眼看就要爆发，就在这关头，经理再一次及时干预。他深晓伽利略关于无限集合的教导，于是告诉巴士司机：完全没问题，在希尔伯特旅馆，任何人都能有房间。他把惊慌失措的前台接待员叫到一边，又给他上了一课。“我们只需要这样做。”他说，“我们让1号房的住客换进2号房，2号房的换到4号房，3号房的去6号房，以此类推。总体而言，n号房的客人应该搬到$2n$号房。这样所有奇数号的房就空了出来，无限星际快车的旅客都能入住。完全没有问题! ”

似乎一切都在经理的掌控之中。但是，假设一艘飞船有某种技术，使得实数轴的连续统[1](continuum)上的每一点都对应于一位乘客，如果这艘船出现了，那么即便是经理也无能为力。每个小数一个人的话，希尔伯特旅馆就会彻底挤爆。在下一小节我们就会知道原因。

康托尔的比较法

当你第一次思考这些问题的时候，它们可能会让你大吃一惊。不过，有一点并不难接受，无限集合的性质在某些方面或许与有限集合不太一样，其中一例便是它们与自身的某些子集有同样的大小。在19世纪，格奥尔格·康托尔(Georg Cantor)比我们走得远得多。他发现并非所有无限集合都拥有同样多的元素，这个发现出人意料，不过你一旦注意到它，就不难体会其中的含义。

康托尔要我们考虑以下这个问题。假设我们有一张无限长的数表L，里面有数a_1，a_2，…，可以把它们看作以小数形式给出的。那么可

以写下一个在L中从未出现过的数a。我们只需要让a的小数点后第一位与a_1的小数点后第一位不同，小数点后第二位与a_2的小数点第二位不同，小数点后第三位与a_3的小数点第三位不同，以此类推。这样，我们就构造出了数a，它与列表中任意一个数都不同。这个结论表面上看起来人畜无害，但它导致了一个直接后果，那便是数表L**绝对不可能**包含**所有**的数，因为L中缺失了数a。由此可知全体**实数**的集合，即所有的小数展开式，不能被写在一个列表里。换句话说，实数集与自然数**不能**像图8里那样建立起一一对应的关系。这条推理链被称为**康托尔对角线论证法**(Cantor’s Diagonal Argument)，为了构造集合L外面的数a，我们想象了L的小数列表形式(如图9)，并用这个阵列的对角线定义了a。

这里有一些微妙之处。我们或许能把缺失的那个数放在L的开头，这样就能轻易绕过这个困难。这创造出一个更大的列表M，它包含引起麻烦的数a。但是，背后的问题并没有消失。我们可以再一次运用康托尔的构造方法，引入一个崭新的数b，而它不存在于这个新列表M当中。我们当然可以像之前一样，无限次地继续加长现有的列表，然而康托尔的结论依然成立：纵使我们能够不断造出含有之前忽略掉的数的列表，但是永远不可能有那么一张表包含每一个实数。

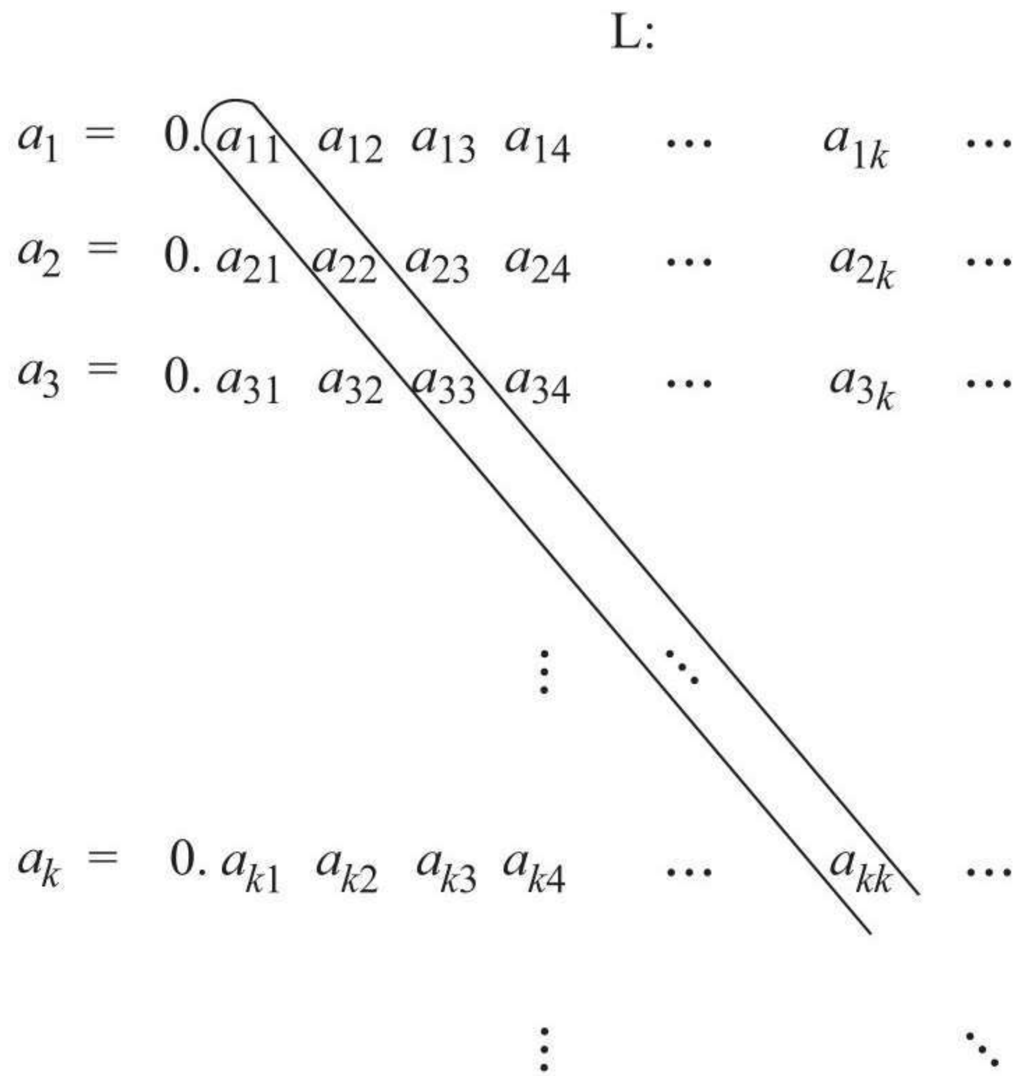


图 9　康托尔对角线论证法

因此，全体实数的集合在某种意义上比所有正整数的集合大。虽说两者都是无穷的，但是不像偶数可以与自然数列表匹配那样，它们没法配对在一起。的确，假定有一张表包含了区间0～1内所有的数，我们将康托尔的对角线法应用在这个列表上，那么缺失的数a也将位于这个范围内。因此，类比于之前，我们同样可以总结出，任何想列出

这个区间的全部元素的尝试都会是徒劳的。我们提起这个事实，是因为很快就要用到它。

其实，很多其他的数集是可以放进一张无穷列表中的，包括希腊人的欧几里得数。在这一背景下，康托尔的结果就显得更加出人意料。这里需要一点点创造力，不过一旦学会几个技巧，就不难证明很多数集是**可数的**(countable)。我们用这个术语来表示集合可以像自然数一样被逐项列举出来。如若不然，则将一个无限集合称作**不可数的**(uncountable)。比如，让我们取全体整数的集合Z，对我们来说，这个集合的自然形态是某种向两个方向无限延伸的列表。不过，我们可以将它重新排列成一行有起点的数：0， 1，-1， 2，-2， 3，-3，…。通过将正整数和各自的负值放在一起，每个整数都会出现在我们造出的列表里——无一遗漏。更令人惊讶的是，我们也可以对有理数做同样的处理。从0开始，接着列出所有可以用不超过1的数来表达的有理数(即1和-1)，然后是只涉及不超过2的数(即 $2, -2, \frac{1}{2}, -\frac{1}{2}$)，再下面是那些只用到不大于3的，等等。用这个方法，所有的分数(正的、负的和0)都可以被安置在一个数列中。于是有理数也构成了一个可数集合。欧几里得数也是一样的，如果从有理数出发取任意次的方根，我们考虑结果得到的数组成的集合，它依然是可数的。我们甚至还可以比这个再进一步：所有代数数(在第6章中第一次出现)——就是普通多项式方程的解——形成一个集合，这个集合原则上可以被排列成一个无穷列表。换句话说，我们可以说出一套系统的方法，将这些数全部列出来。(顺着跟有理数同样的思路，可以证明这个结果。)

我们随意地接受任何可能的小数展开的时候，就打开了通往超越数的大门。那些数超出了能从欧氏几何和普通代数方程产生的数的范围。康托尔的证明告诉我们，超越数是存在的，并且有无穷多个。因为假如它们仅仅组成一个有限的集合，那么它们就可以被放在我们的代数数(非超越数)的表单的开头，这样就产生了一张全体实数的列

表，而我们已知这是不可能的。令人惊异的是，我们发现了这些奇怪的数是存在的，却还没有指认出其中的任何一个! 仅通过互相比较某些无限集合，我们就揭示了这些数的存在。在我们熟悉的代数数和所有小数展开的集合中间，有着巨大的空隙，而超越数正是填充这些空间的数。用一个天文学的比喻来说，超越数就是数的世界中的暗物质。

从有理数到实数，用数学家们的话来说，我们是从一个集合转到了另一个**势**[2](cardinality)更大的集合。如果两个集合的元素可以一对一地匹配起来，那么它们就等势。用康托尔的方法可以证明，任何集合的势都小于由它的所有子集组成的集合的势。这对于有限集合来说是显然的：的确，在第5章，我们已经解释了如果一个集合有n个元素，那么可以构造出2^n个子集。但是，自然数这个无限集合{1，2， 3，…}的所有子集组成的集合S到底有多大呢？这个问题不光本身很有趣，我们得到答案的方式也耐人寻味。结论是S确实是不可数的。

罗素悖论

假设反过来，S是可数的，在这种情况下自然数的子集可以按某种顺序列举出来：A_1，A_2，…。现在，任意一个数n可能是A_n的元素，也可能不是。让我们考虑由所有**不属于**集合A_n的数n所组成的集合A。现在，A是自然数的一个子集(有可能是空集)，于是它应该也在某个时候出现在了之前说的列表里，比方说$A = A_j$。现在出现了一个无法回答的问题：j是A_j的元素吗？如果答案是“是”，那么由A定义方式，我们可以推出j**不是**A的元素。但是$A = A_j$，这自相矛盾。另一种可能的答案是“不是”，j并非A_j的元素，在这种情况下，再一次，通过定义我们推断j是$A=A_j$的一个元素。于是我们又遇到了矛盾。因为矛盾不可避免，我们的原始假设——自然数的子集能被可数地列举，必为假。的确，这个方法成功证明了任何可数无限集合的子集所组成的集合是不可数的。

这种自指式风格的证明是由伯特兰·罗素(Bertrand Russell)引入的，当时的背景稍有不同，他是为了导出罗素悖论(Russell’s Paradox)。罗素将这个方法应用于“所有不属于自己的集合组成的集合”，他问出了以下令人尴尬的问题：这个集合是否是自身的一个元素？再次，“是”会推出“否”，而“否”会得出“是”，这迫使罗素得出结论，这样的集合不可能存在。

19世纪90年代，康托尔自己就从“所有集合组成的集合”这个思想中发现了一个隐含的矛盾。的确，罗素承认他的悖论受到康托尔工作的启发。无论如何，我们从所有这些讨论中总结出的经验是：不能简单地想象数学集合可以随意引入，对于如何描述集合需要设置一些限制条件。从那以后，集合论学家和逻辑学家一直在为这些悖论造成的后果绞尽脑汁。这些困难最令人满意的解决方案，是现在已经成为标准的ZFC**集合论**(包含**选择公理**的策梅罗-弗兰克尔集合理论[3])。

显微镜下的数轴

不同类型的数交织在一起，将整个数轴联结成了一个连续体。当我们观察这些类型的数的分布时，就得到不同的无穷数集。看待这些集合的大小可以有不同的方式。有理数只是一个可数集，但是它密集地填充在数轴上，其分布的方式与整数大不相同。给定任意两个不同的数，a和b，总存在一个有理数在它们之间。这两个数的平均值$c=\frac{a+b}{2}$，当然是一个在它俩之间的数，不过它可能是无理的。但如果c是无理数，我们可以用一个有理数d来近似它，我们可以用有限小数，也就是让d和c直到小数点后很多位拥有相同的数字。如果取$\sqrt{2}$=1.414作例子，我们知道$\sqrt{2}$和1.414相差不到0.001，而每多取一位，就可以保证得到的有理数能更精确地逼近$\sqrt{2}$(平均来说，十倍于之前的精度)。只要前面这些一致的数位足够多，那么这两个数的差别就会很小，于是c和d都会位于a和b之间。小数点后需要的位数取决于一开始的a和b相差

多大，但我们总是能够找到满足条件的有理数d(见图10)。正是出于这个原因，我们说有理数集在数轴上是**稠密的**(dense)。当然，我们还可以用同样的论证，说明从a到d的区间也会被另一个有理数分开。这样，我们就能得出结论，对于任意两个数，无论它们的差别有多小，中间一定存在无穷多个有理数。特别强调，不存在一个叫作最小正分数的东西，因为给定任意正数，在它和0之间总存在一个有理数。

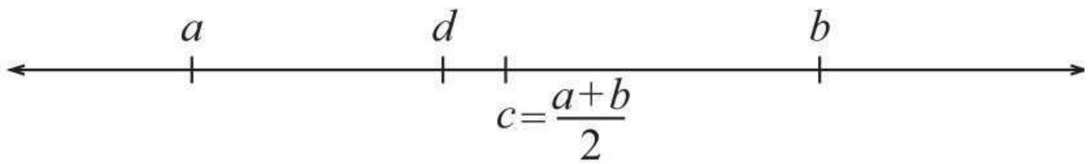


图 10　数轴上任意给定的不同位置都被有理数分隔开

无理数也不甘示弱，它们也构成了一个稠密的集合。在解释这一点之前，让我先指出一旦我们找到了一个无理数，比方说毕达哥拉斯数$\sqrt{2}$，洪水的闸门就打开了，我们立即可以找出无穷多个。当我们让一个有理数加上一个无理数时，结果总是无理数。例如，$\sqrt{2}+7$是无理数。类似地，如果我们将一个无理数乘上一个有理数(非0)，结果会得到另一个无理数。这两个说法都可以用简单的反证法证明。特别是，我们可以找到想要多小就有多小的无理数：$t=\frac{\sqrt{2}}{n}$对于任何自然数n都是无理的，当我们取越来越大的n时，t就会无限接近0。因此我们看到，就像有理数一样，不存在最小的正无理数。所以，也不存在最小的正数这样的东西。

回到给定的数a和b。再一次，令c为它们的平均值。如果c是无理数，我们就得到了要求的数(无理数)。另一方面，如果c是有理数，设$d=c+t$，这里t是上一段我们说到的无理数。根据之前的结论，d也是无理数。如果我们让t里面的n足够大，总可以保证d足够接近两个给定数

a和b的平均值c，从而也位于它们之间。这样，我们看到无理数同样组成了一个稠密集。就像有理数，我们也可以推知数轴上任意两个数之间有无穷多无理数。

因此，有理数集和它的补集——无理数集，在一方面是类似的(它们在数轴上都是稠密的)，在另一方面却又不同(前者是可数的，后者则不是)。

康托尔三分点集

现在，对于有理数和无理数如何织就整个实数轴，我们已经有了更清晰的认识。有理数组成了一个可数集合，但同时也稠密地填充在数轴上。与之相反，康托尔三分点集(Cantor’s middle third set)是单位区间的一个不可数子集，但却是稀疏分布的。它是如下文的构造的产物。

我们从单位区间I开始，也就是从0～1的所有实数，包括0和1。构造康托尔集的第一步是删掉这个区间中间的$\frac{1}{3}$[4]，也就是$\frac{1}{3}$和$\frac{2}{3}$之间的所有数。剩下

的集合包括两个区间：从0～$\frac{1}{3}$，以及从$\frac{2}{3}$～1。第二步，我们将这两个区间的中间$\frac{1}{3}$分别去掉。第三步，我们移除剩余区间各自的中间$\frac{1}{3}$，以此类推。在这个过程中，I内部一直**没有**被移走的点组成了康托尔三分点集——C。



图11　康托尔三分点集第一到四层的形成过程

在这个过程中，当我们从一步前进到下一步的时候，根据定义，剩余小区间的总长度会变成前一个阶段的$\frac{2}{3}$。由此可知，在第n步，剩余区间总长为$(\frac{2}{3})^n$。随着n增大，这个表达式趋向于0。由于康托尔集C是由所有步骤结束后剩下的点组成，可推出C的“长度”，或者叫**测度**(measure)，一定为0。

我们可能会怀疑，在我们倒洗澡水的时候把宝宝也一起倒了出去，或许C里面根本不剩下什么了。三分点集是空的吗？答案是掷地有声的：不! 有无穷多的数留在了C里面。要看到这一点，最好的办法是我们将区间中的数转换成以3为底的“小数”来表示，即三进制，因为整个构造过程就是基于$\frac{1}{3}$的倍数。在三进制小数中，数$\frac{1}{3}$和$\frac{2}{3}$被分别赋予符号0.1和0.2。去掉单位区间的中间$\frac{1}{3}$的时候，我们扔掉了所有的三进制小数以0.1打头的数。实际上，在整个过程中，我们恰好去除了所有那些三进制展开中出现过1的数[5]。C里面正是那些完全由数字0和2组成三进制展开式的数。（比如，$\frac{3}{4}$从这个无穷的剔除过程中幸存了下来，因为在三进制下，它具有循环展开式0.202 020 20…。）

接下来我们将会有一个令人惊喜的发现。取C中任意数c的三进制形式，将每个2都替换成1，我们就得到了单位区间中某个数的二进制展开式。这在C和I的所有数(写成二进制形式)之间建立了一一对应关系。由此可知，C的势与I相等。由于后者是一个不可数集(由康托尔对角线论证可得)，因此康托尔三分点集不仅是无穷的，还是不可数的。

于是我们就有了一个集合C，从某个意义上讲它的大小可以忽略(测度为0)，但是用另一种方式估计，C又是巨大的，因为它与I——因而与整段实数轴——等势。

另外，与稠密集相去甚远，C是**处处不稠密的**(nowhere dense)。回忆一下，当我们说一个像有理数这样的集合是稠密的，我们指的是当我们取一个数a时，在a附近的一个小区间里总能找到有理数，不管这个区间有多小我们都能做得到。我们说a的任意**邻域**(neighbourhood)都包含有理数集的元素。康托尔集拥有完全相反的性质——不属于C的数在数轴上可能从来不会遇上C里的数，前提条件是它们将视野限制于所在位置附近足够狭窄的区域。想要看清这一点，取任一不在C中的数a，于是a的三进制展开式中至少有一个1：a=0.…1…，比如说第n位上的是1。a附近有一个足够小的区间，其中的任意数b的三进制展开式有不止前n位与a的相同，那么所有这些数都**不属于**怪异的集合C，因为它们的三进制展开式也都包含至少一个1。

另一方面，康托尔集的任何数a也并不会感到太孤独，因为当a观察数轴上任何包含它自己的区间J，不论多么小，a总能在身边的邻居中找到同在C里的元素(也有不在C中的数)。我们可以让给定区间J里的一个数b同时属于C，只需规定b的三进制展开式拥有与a足够多相同的位数，但又不包含1。实际上，J包含有不可数个C的成员。

总结一下，康托尔三分点集C拥有可能有的最大数量的元素，当C的成员左右看去的时候，它们的兄弟姐妹在周围到处可见。然而，对

于不属于C的数，C就像不存在一样。在它们的邻域内，看不到一个C中元素的身影，而集合C本身的测度也为0。对它们来说，C几乎什么也不是。

丢番图方程

数轴上的某些重要的集合可以用方程的语言来描述。组成可数集的有理数，是简单的线性方程的解：分数$\frac{b}{a}$是方程$ax-b=0$（a和b是整数）的解$a\neq 0$。像$\sqrt{2}$一样不能由这种方式产生的数称作无理数，并组成了一个不可数集，它们不像有理数那样能跟自然数配起对来。无理数中又有超越数，即便我们允许以上这类方程中出现x的高次项，也永远得不到超越数。人们已经知道，π 是一个超越数，但$\sqrt{2}$不是，因为它是方程$x^2-1= 0$的解。这种方法通过不同类型方程的解来定义数的种类。

不过，当我们往反方向走时，产生了一条富有趣味的研究路线。我们要求不仅方程的系数是整数，并且解也得是整数。这里举一个经典的例子。

一个装着蜘蛛和甲虫的盒子里有46条腿，问两种生物各有多少只？这个小小的谜题可以用试错法轻松解决，但是仔细观察下面两点，我们可以学到些东西。第一，它可以被一个方程来表示$6b+8s=46$。第二，我们只对这个方程的某些种类的解感兴趣，即甲虫(b)和蜘蛛(s)的数量都是整数的解。一般来说，当我们只限定于搜寻特殊类型的解时，相应的方程组就被称作**丢番图方程**(Diophantine equation)。通常情况下我们要找的是整数或是有理数解。

要求解本例中这样的线性丢番图方程，有一个简单的方法。首先，将方程除以系数的最大公因数。在这个例子中，系数是6和8，因此它们的最大公因数是2。消去这个公因数2之后，我们得到一个等价

的方程——也就是说一个有同样解的方程：$3b+4s=23$。如果做完这次除法之后，右侧不再是整数，那就表明该方程没有整数解，我们可以就此罢手。下一步骤是取系数当中的一个(通常用最小的那个，因为这样最简单)，并将方程用它的倍数来表达。这里最小的系数是3，我们的方程可以写成$3b+3s+s=(3\times 7)+2$。整理后我们得到$s=(3\times 7)-3b-3s+2$。这样做的目的在于揭示出s有$3t+2$这样的形式，这里t是某个整数。将$s=3t+2$代入我们的方程，整理出b的表达式。我们得到

$$3b+4(3t+2)=23\rightarrow 3b=15-12t\rightarrow b=5-4t.$$

现在，我们有了丢番图方程完备的整数解：$b=5-4t$，$s=3t+2$。t选择任意一个整数值都给出一组解，而所有整数解都具有这样的形式。

当然，我们的原始问题还有附加的限制条件，即b和s两者都大于等于0，因为不存在负数个蜘蛛或者甲虫。因此，只有两种可行的t值，$t=0$或$t=1$。这给了我们两组可能的解，5只甲虫和2只蜘蛛，或者1只甲虫和5只蜘蛛。假如我们对原题咬文嚼字，认为两种生物都必须有不止一只[6]，那么我们便得到传统的答案：5只甲虫和2只蜘蛛。

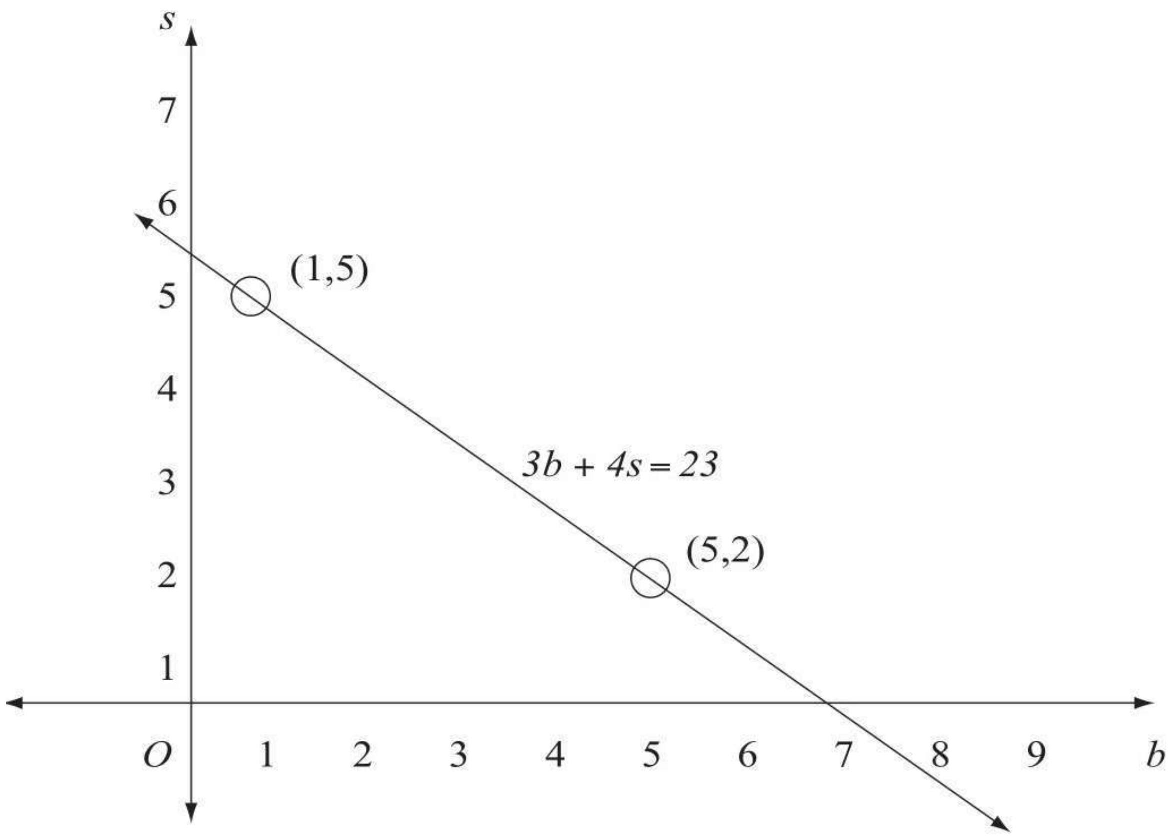


图 12　线性方程图像直线上的栅格点

这种类型的问题称作**线性的**(linear)，因为它所涉及方程的图像由排成一条直线的无穷多个点组成。丢番图问题则是要找到这条直线上的**栅格点**(lattice points)，也就是那些两个坐标值都是整数的点。或者，如果我们只允许正数解，则只能是在第一象限里面的栅格点。

然而，一旦我们允许方程中出现平方和更高次方，对应的问题的性质会变得更加多样化和有趣。在这种类型中的一个经典问题上我们已经获得了完整的解，那就是找出所有的**勾股数**[7](Pythagorean triple)：满足$a^2+b^2=c^2$的正整数a，b，c。当然，得名勾股数是因为你可以用这三个正数作为边长画出一个直角三角形。典型的例子是(3，

4， 5)三角形。给定任意勾股数，我们只需要将它们乘上任一正整数，就能得到更多的勾股数，因为勾股方程依然成立。例如，我们可以将刚才的例子翻倍，得到(6， 8， 10)这组数。不过，这给出的是一个相似三角形，它和原来那个有一模一样的比例关系，因为改变的只是整体尺寸而不是形状。给定第一个三角形，我们只需用原来单位的一半来测量边长，就可以使三角形的尺寸数值变为之前的两倍，也就得到了第二组勾股数。但是，也有本质上不一样的勾股数，比如(5， 12， 13)和(65， 72， 97)。

所以，要描述所有的勾股数，只需要描述最大公因数为1的所有三元素组(a， b， c)就行了，因为一切其他的勾股数都仅仅是这些的放大版。方案如下。取任意一对互质的正整数m和n，其中一个为偶数。假设m表示两者中大的那个数。构造三元数$a=2mn$， $b=m^2-n^2$及$c=m^2+n^2$。那么这三个数a，b，c就给出一组勾股数(这里用代数可以轻松验证)，并且这三个数没有公因数(也不难检验)。上面的三个例子就可以这样得到。在第一个例子里取m=2和n=1，第二个里取m=3和n=2，而最后一种情况我们有m=9，n=4。要检验反过来的命题需要更多努力：任何勾股数都可以用这种方法来生成，只需取合适的m和n的值；另外，这样的表示法还是唯一的，于是两对不同的(m，n)不可能产生同样的勾股数(a，b，c)。

相应的立方和更高次的方程却没有解：对于任何指数$n\geqslant 3$，没有正整数组x，y和z满足$x^n+y^n=z^n$。这就是著名的**费马大定理**(Fermat’s Last Theorem)。未来它可能被称作怀尔斯定理，因为它是由安德鲁·怀尔斯爵士[8](Sir Andrew Wiles)在20世纪90年代最终证明的。甚至是在立方的情况下，这也是一个非常困难的问题，这是由欧拉首先解决的。不过，可以相对简单地证明两个四次方数之和不可能是一个平方数(因此当然不可能是一个四次方)。并且这就足以将原来的问题简化为指数n是一个素数p的情况(意思是如果我们对所有素指数的情况解

决了这个问题，便可以立即得到一般的结论）。确实，在19世纪的时候，指数n为所谓的**正则素数**(regular prime)的情形得到了证明。然而，要实现问题的完全解决，还得等到怀尔斯解答了一个叫作谷山-志村猜想[9](Shimura-Taniyama Conjecture)的深刻问题之后。

不过，被研究最多的丢番图方程是**佩尔方程**(Pell equation)，$x^2-ny^2 = 1$，这里n是一个正整数，但不是平方数。人们在很早以前就已经认识到了它的重要性，早至公元前400年，希腊和印度似乎就有人研究过它了，因为这个方程的解给出了$\sqrt{n}$的很好的有理数近似，即$\frac{x}{y}$。例如：当n= 2时，方程有一组解，数对x=577和y=408，并且$(\frac{x}{y})^2$。这个方程与一个希腊人称为**辗转相减**(anthyphairesis)的几何过程有关。在这个过程里，我们从两条线段出发，反复地从长的线段中截去短的那段，有点类似于欧几里得算法，不过是应用在有连续取值的长度上。实际上，第3章中提到的阿基米德牛群问题恰好也会导出一个佩尔方程。

公元150年前后，丢番图自己研究了佩尔方程的一些版本。不过，是伟大的印度数学家婆罗摩笈多(Brahmagupta)解出了这个方程，他的方法又被婆什迦罗(Bhaskara II)所改进，后者揭示了如何从已有的种子解出发创造新的解。在欧洲，是费马敦促数学家们注意佩尔方程，而完整的理论则要归功于著名的法国数学家约瑟夫-路易·拉格朗日(Joseph-Louis Lagrange)（英语名称“佩尔”是历史上的一个偶然事件）。一般的解法基于$\sqrt{n}$的连分数展开。

斐波那契和连分数

回忆一下斐波那契发现的数列，1， 1， 2， 3，5， 8， 13，21，…，我们在第5章中进行了介绍。从这个数列中取一对相邻的数，将它们的比写成1加上一个分数。现在，如果我们“埃及化”这个分

数，也就是反复地让分数部分的分子分母同除以分子，就会出现一个引人注目的规律。比方说

$$\frac{13}{8}=1+\frac{5}{8}=1+\frac{1}{1+\frac{3}{5}}=1+\frac{1}{1+\frac{1}{1+\frac{2}{3}}}=1+\frac{1}{1+\frac{1}{1+\frac{1}{1+\frac{1}{1+\frac{1}{1}}}}}.$$

我们得到了一个仅含有1的多层分数，并且随着计算持续进行，选出的这对相邻数之前的每对斐波那契数的比值会依次出现。这样的现象每次都会发生：正是由于这些数的定义，每个斐波那契数都比下一个的两倍要小，因而相除之后会得到商1，余数则是前一个斐波那契数。你也许想起来，相邻斐波那契数的比值趋近于黄金分割比——τ，这说明τ是仅由1构成的连分数的极限值。

正如第5章所介绍的，如果一个值产生于一个无穷多次重复的过程，那么它也许能用一个基于此过程的方程来表示。假设我们将1组成的无穷分数塔的值设为a，我们看到a满足$a=1+\frac{1}{a}$这一关系，因为位于分数第一层分号下方的，正是另一整个a。从这里，可以看出a满足二次方程$a^2=a+1$，它的正根是$\tau=\frac{1+\sqrt{5}}{2}$。

从这个过程中产生的连分数本身就有重要意义。当用有理数来近似任意一个无理数y的时候，我们会自然地使用y的小数表示。这非常适用于一般的计算，但是在数学上，依附于一个特定的底数却不太自然。本质问题在于，我们能多好地用分母相对较小的分数来逼近y。有没有一种办法，能找到一系列的分数，既能高精度逼近y同时又保持较

小的分母，在这两个矛盾的要求中取得最佳平衡？答案就在于一个数的连分数表达形式，并通过越来越多层的截断实现这一目的。

由于我们在表达式中使用的多层分号，连分数看起来十分别扭。但是，我们可以轻易避开书写所有这些分号的不便——因为每个分子都是1，为了指明我们所说的是哪个连分数，只需要记录相除后的那些商。比如，分数$\frac{25}{91}$的表达式可以如下推演：

$$\frac{25}{91}=0+\cfrac{1}{3+\cfrac{16}{25}}=0+\cfrac{1}{3+\cfrac{1}{1+\cfrac{9}{16}}}=\cdots.$$

最终产生的连分数可以用列表[0， 3， 1， 1，1， 3， 2]来代表。正如我们已经看到的，黄金分割比的连分数表达为[1， 1， 1，1，…]。用一个类似于循环小数的记号，我们写作$\tau=[\bar{1}]$。在$\frac{25}{91}$的连分数中，第一个商3是从91 = 3× 25+16得来的，这也是对(91， 25)应用欧几里得算法得到的第一行等式。事实上，正是由于这个原因，作用于两个数的欧几里得算法中的每一行都对应于连分数中的一行。特别是，从一个**既约分数**[10](reduced fraction)(分子分母互质)出发，在计算对应的连分数的过程中出现的每一个分数，都适用于同样的规律。

黄金分割比所提供的这个特殊例子打开了一条新的思路，即我们也许能表示其他的无理数，当然不是用有限的连分数(它们自己显然是有理数)，而是用无穷连分数。但是，怎样才能产生一个数a的连分数

呢？为了看清这个过程，读者朋友可能需要容许我玩一个小小的代数上的把戏，下面就是做法。

计算一个数a的连分数a = [a_0， a_1， a_2…]有两步。数a_0是a的整数部分，用$a_0=\lfloor a \rfloor$[11]表示。(例如，π =3.1415926的整数部分由$\lfloor \pi \rfloor = 3$给出。)一般来说，$a_n=\lfloor r_n \rfloor$，即r_n的整数部分，这里余数项r_n是由r_0=a，$r_n=\frac{1}{r_{n-1}-a_{n-1}}$递归定义的。将这个方法应用于a=$\sqrt{2}$，并且使用分母有理化的代数技巧(有些读者对此可能已经很熟悉了)，考虑到$\lfloor \sqrt{2} \rfloor = 1$，我们得到

$$a = r_0 = \sqrt{2} = 1 + (\sqrt{2}-1), \text{ 于是 } a_0 = 1 ;$$

$$r_1 = \frac{1}{r_0 - a_0} = \frac{1}{\sqrt{2}-1} = \frac{\sqrt{2}+1}{(\sqrt{2}-1)(\sqrt{2}+1)} = \sqrt{2}+1, \ a_1 = \lfloor r_1 \rfloor = 2;$$

接下来我们得到

$$r_2 = \frac{1}{r_1 - a_1} = \frac{1}{\sqrt{2}+1-2} = \frac{1}{\sqrt{2}-1} = \frac{\sqrt{2}+1}{(\sqrt{2}-1)(\sqrt{2}+1)} = \sqrt{2}+1 .$$

$$\text{因此} r_1 = r_2 = \cdots = \sqrt{2}+1, \ a_1 = a_2 = \cdots = 2, \text{ 所以} \sqrt{2} = [1, \bar{2}] \text{。}$$

事实上，拥有循环的连分数表达式的数包括有理数(即有限的连分数表达式)，以及二次方程的解。比如说τ ，上面我们已经看到了τ 是方程x^2=x+1的一个解；还有$\sqrt{2}=[1,\bar{2}]$，它满足x^2=2，其他一些例子包含的循环则更令人难以捉摸，$\sqrt{3}=[1,\overline{1,2}]$，[illegible]。不过，无理数平方根的连分数展开式有着非常特殊且引人注目的一面。这些展开式以整数r打头，接着它们的

循环部分是一串回文数列(palindromic sequence)(数列中的数反着读与正着读一样)，再后面跟着$2r$。在前面所有的例子中都能观察到这一点：比方说$\sqrt{28}$，我们看到$r=5$，展开式的回文部分为3，2，3，后面跟着$2r=10$。对于$\sqrt{2}$和$\sqrt{17}$，回文部分是空的，但规律依然存在，只不过是以一种简单的形式出现。可以证明，一个数的连分数表达式是唯一的——两个不同的连分数有不同的值。

在用有理数对无理数的逼近中，连分数的重要性在于所谓的**收敛子**[12](convergent)，即原始数的有理数近似，它们是通过截取某一层以前的连分数表达式，再求出相应的有理数来得到的。它们代表了对原始数的最优近似，也就是说任何更好的近似的分母都将比收敛子的分母大。黄金分割比的收敛子是斐波那契比值。由于τ的连分数表达式中每一项都是1，这些比值收敛的速度被尽可能地延缓了。出于这个原因，没有比τ更难用有理数逼近的数了，而斐波那契比值是你能取得的最佳结果。

如果一个连分数的某个收敛子的分母为q，那么这个近似值总是在距离真实数值$\frac{1}{\sqrt{5}q^2}$的范围之内，并且连分数的收敛子轮流低估和高估被逼近的值。然而，在有理数逼近这个问题上，像τ和$\sqrt{2}$这样的欧几里得数总是最难的。有一些特殊的超越数，它们的性质似乎与有理数的世界离得不能再远了，但依然可以被很好地逼近，它们的收敛子以极快的速度接近目标。

上一节末尾我们提到了佩尔方程，$x^2-ny^2=1$，它与这里的联系在于，如果要找方程的解(x, y)，使x取最小的可能正值，那么这样的解存在，并且可以在$\sqrt{n}$的连分数的收敛子中找到。例如，当$n=7$时，$\sqrt{7}$的收敛子数列的前几项为$\frac{2}{1}, \frac{3}{1}, \frac{5}{2}, \frac{8}{3}$。正是$x=8$，$y=3$提供了佩尔方程$x^2-7y^2=1$的最小的解，也就是所谓的**基本解**(fundamental solution)。不过，有时候基本解并不会很早就在展开式中现身：比如，$x^2-29y^2=1$的最小

正解为x=9801和y=1820。当然，一旦这个基本解(x，y)被确定，其他所有的解都可以通过以下途径得到：求连续的次方$(x+y\sqrt{n})^k$(k=1,2,3…)，得到展开括号后的表达式，再分别提取有理和无理部分的系数。用这个方法，佩尔方程的完整解集就可以通过$\sqrt{n}$的连分数表达式求得。

[1] 通常称实数集(直线上点的集合)为连续统。

[2] 有时也称基数。

[3] 恩斯特·策梅罗(Ernst Friedrich Ferdinand Zermelo)，德国数学家。亚伯拉罕·弗兰克尔(Abraham Fraenkel)，以色列数学家。

[4] 删掉的是开集，即($\frac{1}{3}$，$\frac{2}{3}$)。

[5] 在三进制中，0.1可用0.0222…表示。

[6] 在英文原题中，“蜘蛛”和“甲虫”都使用了复数形式的名词，即它们各自不止一只。

[7] 勾股数或称商高数或毕氏数。

[8] 安德鲁·怀尔斯爵士(Sir Andrew Wiles)，英国数学家，居于美国。

[9] 谷山丰和志村五郎，均为日本数学家。

[10] 又称最简分数。

[11] 该符号表示向下取整。

[12] 又称渐进分数。

08 并非我们熟知的数 Numbers but Not As We Know Them

实数和复数

构造复数的过程比构造实数容易得多，实施起来也顺利得多。创造实数的第一个步骤是发展出有理数，这时候我们得解释到底什么是分数。一个分数，譬如说$\frac{2}{3}$，其实仅是一对整数，但是我们用这个既熟悉又怪异的方式来表达。分数的思想不难理解，但是相应的算术操作却需要一些努力才能掌握。在这个过程中，你的老师顺便提了一下，像$\frac{2}{3}, \frac{4}{6}, \frac{6}{9}$这些分数都是“相等的”——虽然它们是不同的数对，但在切馅饼时，它们代表了相同大小的分块。这接受起来并不难，不过它使得我们注意到一个事实，即一个有理数实际上是无穷多等价的分数的集合，其中每个元素都能用一对整数来表示。这听起来很吓人，或许我们不乐意朝这方面想太多，因为想象一下要操作整数对的无限集合，就让人感到不舒服。我们有一根救命稻草，那就是任何分数都有唯一一个简化形式，也就是当分子和分母互质的时候。从一开始的分数中消去所有公因数，你就能得到这个最简形式。一旦你熟悉了分数的性质和使用它们的规则，就应该不会出错了。不过仔细观察一下，你会发现当自己做加法的时候，你已经隐含地在摆弄整数对的无限集合了。

然而，当我们进一步努力想要弄清楚实数到底是什么的时候，事情变得更糟了。让我们从毕达哥拉斯的问题说起。他发现没有一个分数等于$\sqrt{2}$，因此我们可以引入一个新的符号r，并使其拥有$r^2=2$的性质。这样从有理数和这个新的数r出发，我们就得到了一个新的数域(field)。这之所以可行，原因在于所有形如$a+br$的数——其中a和b是有理数——都遵循代数的一般规则。我们甚至可以做除法，因为这种

数的倒数也保留了同样的形式。通过叫作分母有理化的炫酷代数小技巧，我们就能看出这一点。

新的数r和$-r$组成了方程$x^2=2$的两个解，那么$x^2=3$呢？我们似乎需要再增加一个新的数，才能求解这个方程，因为很容易验证没有形如$a+br$的数会平方得3。（这里一个简单的反证法就够了：假设$(a+br)^2=3$，这使你推出$\sqrt{2}$和$\sqrt{3}$中至少有一个是有理数的错误结论。）

我们会很想把这些关于特定方程的烦恼都丢在一边，直接声称我们已经知道实数是什么了——它们是所有可能的小数展开式组成的集合，包括正的和负的。这我们已经很熟悉了，在实际应用中大家也知道如何运用它们，因此我们觉得自己的位置是坚实的——至少在我们开始问一些很基本的问题之前。数的主要特点在于你可以加、减、乘和除。但是，举例来说，你怎样才能将两个无限不循环小数相乘呢？我们指望小数的长度有限，然后你就可以“从最右端开始”，但对于无限小数来说没有这样的东西。其实这是可以做到的，不过从理论和实际操作上讲它都很复杂。如果你解释如何加和乘都非常困难的话，这个数的系统似乎就不够令人满意。

还有其他的小陷阱。当你将$\frac{1}{3}$乘以3时，答案为1。当你将0.333…乘以3，答案当然是0.999…。两个不同的小数展开式确实能表示同一个数：1.000…=0.999…。事实上，任何一个有限小数都是这样的，比如0.375=0.374 999…。正因为我们看到两个不同的小数展开可以等于一样的数，所以，我们声称小数和实数是同一个东西严格来说是不太对的。另外，如果我们换一个进制底数，那么拥有不唯一小数展开的数会跟之前不一样，这又增加了额外的复杂性。倘若我们用小数来定义实数，我们就使得这一构造依赖于一个随意做出的选择(十进制)。假如我们用二进制来进行同样的构造，得到的“实数”集还会是相同的吗？以及我们的“相同或不同”是什么意义上的呢？

你也许会觉得上面所提出的基本问题耐人寻味，或者你会对我们的自省感到不耐烦。毕竟，之前所有的航程都是一帆风顺的，我们似乎是在自找麻烦。但有一点是无法忽视的。数学家们认为，任何时候我们引入新的数学对象，重点是要从已知对象出发再构造它们，就像分数可以被看作一对普通整数。这样，我们能够仔细地定义新推广的系统所遵循的规则，从而了解自己所处的位置。倘若我们完全忽视基础，日后它便会出来找麻烦。例如，微积分学脱胎于对运动的研究，它发展得极快，并且获得了光辉的成就，比如预测行星的轨道。然而，像对待有限事物一样处理无限事物，有时候能赋予我们惊人的洞察力，有时候却完全没有意义。将数学系统建立在坚实的基础上，我们就能学会如何分辨真假。在实践中，数学家们经常沉迷于“形式化”(formal)的操作，这是为了能看清远方的海面上是否会浮现出崭新的定理。要是结果值得注意，我们就可以通过回溯基本概念和引用已经恰当地建立起来的结果，来严格地证明它。

这就是为什么尤利乌斯·戴德金(Julius Dedekind)要不辞辛苦形式化地构造实数系。现在，我们将他的思想称为实数轴的**戴德金分割**(Dedekind cuts)。不过，对于无理数存在性导致的两难问题，第一个成功提出解决方案的数学家是尼多斯的欧多克索斯(Eudoxus of Cnidus，他活跃于公元前380年)。借助于他所著的《比例论》(Theory of Proportions)，阿基米德使用所谓的穷竭法(Method of Exhaustion)严格地推导出了弯曲形状的面积和体积，而这比微积分的发明早了大约1900年。

数的最后一块拼图——虚数单位

i的平方为-1，这个新符号的引入是一项巨大的成功，因为它一次性地解决了很多问题，不仅是为方程提供了一个解，同时使得**所有**多项式方程都可解，还帮忙解决了不少其他方面的问题。对于任意负数

——$-r$，我们显然有两个平方根。$\pm i\sqrt{r}$这两个数的平方都是$-r$，这一方面是由于$i^2=-1$这项性质，另一方面是因为我们假设，就像普通的算术一样，乘法**可交换**(commute)，这意味着对于任意两个数z和w，$zw=wz$。如果我们坚持要求复数$a+bi$的系统应该包含实数(对应于$b=0$的情况)的性质，并且所有正常的代数规则也都还适用，我们不会遇到任何困难，反而会发现很多令人愉快的惊喜。复数的集合——表示为C——是一个“域”，域除了一系列其他性质以外，还保证了除法是能进行的。但是，要看出这是怎么实现的，我们最好离开实数轴这条单向轨道，去看看二维世界的样子。

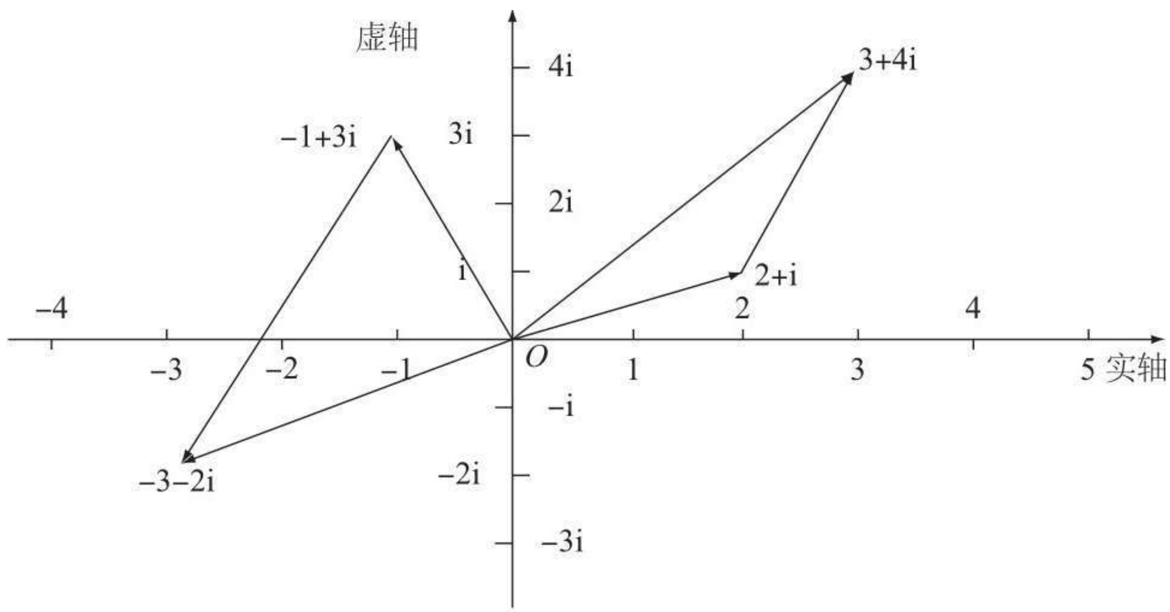


图 13　复数的相加即有向线段的相加

复数的算术可以在**复平面**(complex plane)内清楚地表示。我们将复数$a+bi$看作坐标平面内的点$(a,\ b)$。当我们将两个复数$z=(a,\ b)$和$w=(c,\ d)$相加时，我们只是将它们的第一和第二个元素分别相加，这给我们$z+w=(a+c,\ b+d)$。如果我们使用符号i，那么举个例子，我们就有$(2+i)+(1+3i)=3+4i$。

这对应于平面上的**向量和**(vector addition)，也就是有向线段(**向量**)首尾相加在一起(见图13)。在这个例子里，我们从坐标为(0，0)的**原点**(origin)开始到点(2， 1)结束，画下第一个箭头。要加上(1， 3)代表的数，我们从点(2， 1)开始再画一个箭头，表示在水平方向(即**实轴**的方向)向右移动1个单位，以及在竖直方向(即**虚轴**的方向)向上移动3个单位，最终到达坐标为(3， 4)的点。用几乎同样的方法，我们可以通过实部和虚部分别相减来定义复数的减法。因此例如，(11+7i)-(2+5i) = 9+2i。这可以看作从向量(11， 7)开始，减去向量(2， 5)，在点(9， 2)结束。

乘法就是另外一回事了。形式上说很容易：我们通过把括号拆开将两个复数相乘，记住$i^2=-1$。假设**乘法分配律**(Distributive Law)仍然成立，这让我们能用通常的方法打开括号，那么乘法如下：

$$(a+bi)(c+di)=a(c+di)+bi(c+di)$$

$$=ac+adi+bci+bdi^2=(ac-bd)+(ad+bc)i.$$

另一方面，除法可以用**复共轭**(complex conjugate)的方法来计算。一般地，$z=a+bi$的共轭是$a-bi$，记作$\overline{z}$。换句话说，$\overline{z}$是z相对于实轴的镜像。乘法运算$z\overline{z}$给出a^2+b^2，这是一个实数，因为我们发现虚部为0。这等于从原点到z的距离(记作$|z|$)的平方。用符号写出来就是$z\overline{z}=|z|^2$。现在，我们就可以操作复数除法了，只需要分子分母同乘除数的共轭，这样便能使除数变成一个纯实数。这可以类比于分母有理化——从分号下面消去平方根的标准技巧，当时我们用它计算了$\sqrt{2}$的连分数。例如，

$$\frac{15+16i}{2+3i}=\frac{(15+16i)(2-3i)}{(2+3i)(2-3i)}=\frac{30-45i+32i-48i^2}{2^2-6i+6i-3^2i^2}$$

$$=\frac{(30+48)-(45-32)i}{2^2+3^2}=\frac{78-13i}{13}=6-i.$$

我们可以用一般化的复数，而不是具体的数值，来将复数相除的结果表示成普适的形式，它由两个数的实部和虚部构成，就像我们在复数乘法中做的一样。不过，只要理解了这一技巧，我们就不一定非要推出并记住最后的公式了。

我们如果把坐标系从普通的直角坐标转换成**极坐标**(polar coordinates)，就会发现乘法有了一种几何解释。在这个系统中，一个点z依然由一个有序数对所确定，我们将其写作(r，θ)。数r是从原点O，在这里叫作**极点**(pole)，到我们的点z的**距离**。因此r是一个非负的量，所有具有相同r值的点形成一个圆心在极点、半径为r的圆。我们用第二个坐标θ来表示z在这个圆上的位置，θ是从实轴到Oz这条线逆时针方向走过的**角度**。数r称为z的**模**(modulus，复数形式为moduli)，而角度θ称作z的**辐角**(argument)。

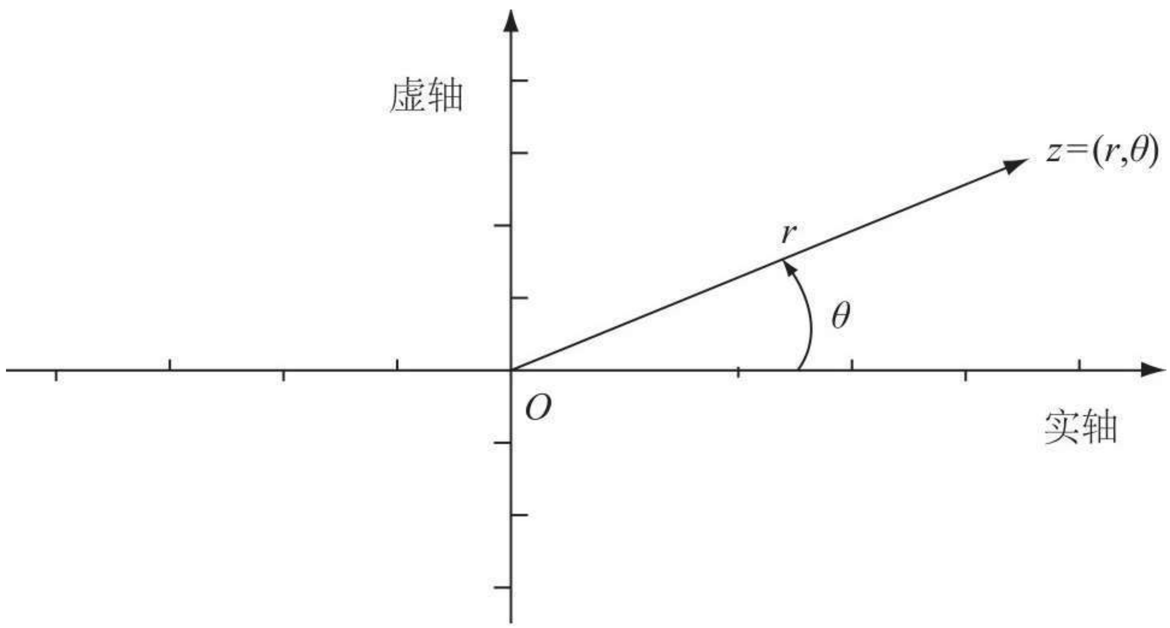


图 14　一个复数在极坐标下的位置

假设现在我们有两个复数，z和w，它们的极坐标分别为(r_1，θ_1)和(r_2，θ_2)。我们发现，它们的积zw的极坐标有一个简单美妙的形式。组合的规则甚至可以用日常语言清晰地表述出来：积zw的模即z和w的模的乘积，而zw的辐角为z和w的辐角之**和**。用符号表示，zw的极坐标为(r_1r_2，$\theta_1+\theta_2$)。实数的乘法包含在这个更一般的规则里：比如，一个正实数r拥有极坐标(r，0)。如果我们乘上另一个数(s，0)，结果是意料之中的(rs，0)，对应于实数rs。

这个表示方式能够更充分地体现复数乘法的特点。复数单位i的极坐标是(1，90°)。通常，在这些情况下，我们并不用度数来度量角，而是用自然的数学单位**弧度**(radian)：一个圆有2π弧度，因而转动一弧度相当于沿着中心在原点的单位圆的周长移动一个单位。一弧度大约是57.3°。假设我们现在取任意复数$z=(r，\theta)$，乘以i =(1，90°)，我们发现zi=(r，θ+90°)。也就是说，**乘以i相当于绕着复平**

面的中心旋转一个直角。再换个说法，直角，这个最基本的几何思想，可以用一个数来表示。

的确，若是将复平面的一个给定区域内的所有点加上或乘以一个复数z，这一效果可以用几何方法来表示。想象平面内任意一个你喜欢的区域，如果给区域内的每一点都**加上**z，我们就只是将每个点都往同一个方向移动相同的距离，这个方向和距离是由z代表的箭头——或者我们经常说的向量——来决定的。也就是说，我们将这个区域**平移**(translate)到平面上的另一个位置，而它的形状、大小和姿态都保持不变，这里姿态不变是指该区域没有经过任何旋转或反射。但是，将你的区域中每个点都乘以$z=(r, \theta)$则有两个效果，一个由r引起，另一个由θ造成。区域内每个点的模都增大r倍，因此该区域的所有尺寸也都增大了r倍(因而它的面积乘了因数r^2)。当然，如果$r<1$，那么我们最好把这个“扩张”描述成收缩，因为新的区域会比原来的小。不过，区域将保持它的形状——例如，一个三角形会被映射为一个相似的三角形，它的各个角和以前一样大。θ的作用就像我们上面已经解释过的，是将区域沿逆时针方向绕极点转过角度θ。那么，将你的区域中所有点都乘上z的总效果是扩展区域，并绕极点旋转。新的区域将和之前的有同样的形状，但取决于r的大小，会有不同的尺寸，同时将会有一个不同的姿态，这是由旋转角θ决定的。

其他结果

复数的极坐标版本尤其适合进行幂和方根运算，因为要求$z=(r, \theta)$的n次方，我们只需要求模的同样次方，并将θ自己相加n次，从而得到$z^n=(r^n, n\theta)$。同样的公式也适用于分数幂和负幂次，除法也可以用极坐标来理解。与实数一样，除以一个复数z意味着乘上它的倒数$w=\frac{1}{z}$，但w是什么数呢？给定$z=(r, \theta)$，w具有性质$zw=(1, 0)$，即相乘为1的那个数。这显示我们必须取$w=(\frac{1}{r}, -\theta)$，因为这样我们才得到要求的[illegible]。在

直角坐标形式中，我们需要使用复共轭，极坐标里则给出了另一种方法。

复数有极多的应用，甚至是在很基础的层次上。直角坐标和极坐标的相互转换将三角函数引入了进来，这种应用方式既令人惊讶又有很多优点。例如，推导重要的三角恒等式是一道标准的学生习题，而在用了极坐标后这些等式是十分自然的结论，取任意单位模(即$r=1$)的复数，用直角坐标和极坐标分别计算它的某次幂，令这两种形式的答案相等，这就给出了一个三角方程。

由基本的三角学可知，极坐标为$(1，\theta)$的点的直角坐标是$(\cos\theta, \sin\theta)$。如果我们现在将两个这样的复数$z=\cos\theta+i\sin\theta$ 和$w=\cos\phi+i\sin\phi$ 在直角坐标中相乘，可得：

$$zw=(\cos\theta\cos\phi-\sin\theta\sin\phi)+i(\cos\theta\sin\phi+\sin\theta\cos\phi).$$

同样的乘法在极坐标中给出：

$$zw=(1, \theta)(1, \varphi)=(1, \theta+\varphi)=\cos(\theta+\varphi)+i\sin(\theta+\varphi).$$

对比该乘积的两个版本的实部和虚部，就可以轻松得到三角学中标准的和角公式：

$$\cos(\theta+\varphi)=\cos\theta\cos\varphi-\sin\theta\sin\varphi,$$

$$\sin(\theta+\varphi)=\cos\theta\sin\varphi+\sin\theta\cos\varphi.$$

或者，极坐标形式的复乘法可以由这些三角公式推导出。实际上，我们在这里未经证明就给出了极坐标形式下的乘法，它通常是将

三角公式应用于直角坐标形式来推导出的。

现在，随着复数的使用，指数函数——或者说幂函数与看起来无关的三角函数之间的联系浮现了出来，更多的结果可以因此轻松得到。假如我们没有跨入由-1的平方根所提供的传送门，我们也许可以窥见这一联系，但却不能理解它。分别取指数函数中被称为偶的和奇的部分，就产生了所谓的**双曲函数**(hyperbolic function)。对于每个三角恒等式，除了符号上可能不同，都相应存在一个等价的由双曲函数表达的形式。在任何具体例子中，这都可以轻松验证。问题是为什么会发生这一现象？一类函数的行为为何会如此紧密地反映在另一类中，而后者来自于完全不同的定义、具有完全不同的性质？解开这一谜团的关键在于公式$e^{i\theta}=\cos\theta+i\sin\theta$，它表明指数函数和三角函数其实是紧密相连的，但这只能通过使用虚单位i做到。一旦发现了这一点(它令人惊讶，一点也不明显)，用这个公式提供的两种可互换的表示方式计算，再令实部和虚部分别相等，我们就可以清楚地看到之前描述的那些结果都是顺理成章的。但要是没有这个公式，所有这些就依然是个谜。

复数和矩阵

乘以i代表绕复平面的中心转动一个直角，让我们来仔细研究一下这个事实所导致的一些结果。若$z=x+iy$，展开括号和对乘法重新排序，可得$i(x+iy)=-y+ix$，因此点(x, y)经过这个旋转变成了$(-y, x)$，见图15。这样，乘以i可以看作在平面上对点操作。这一操作有一个特殊性质，即对于任意两点z和w，以及任意实数a，我们有$i(z+w)=iz+iw$，和$i(aw)=a(iw)$。

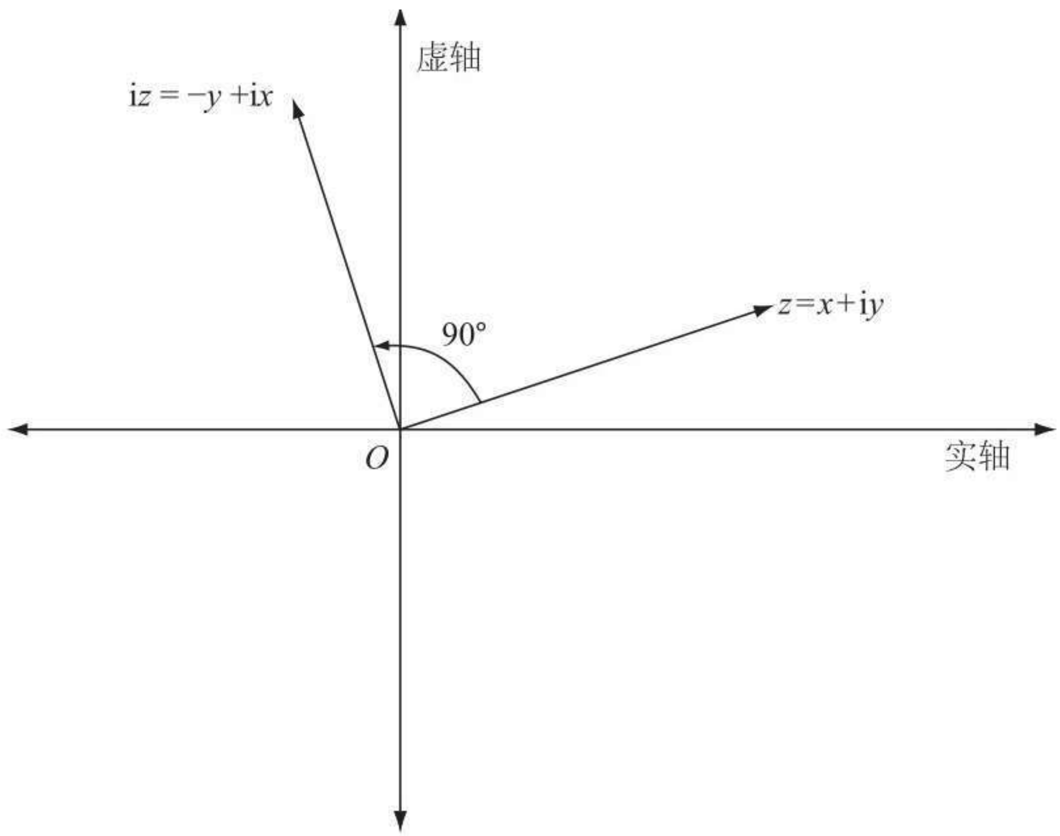


图15　乘以i将一个复数旋转一个直角

另外，如果我们将一个实数a乘上一个复数(x+iy)，我们得a(x+iy) = ax+i(ay)。用复平面上的点来说，我们将(x， y)移动到了(ax，ay)，或者用另一种方法写就是a(x， y) = (ax， ay)。

具有这两条性质的一类运算称为**线性的**(linear)，它们在数学的所有领域都极其重要。这里，我只希望提醒你注意一个事实，那就是这样一个运算L的效果完全由它在两个点(1， 0)和(0， 1)上的行为决定。因为让我们假设L(1， 0) = (a， b)和L(0， 1) = (c， d)，那么对于任何点(x， y)，我们有(x， y) = x(1， 0)+y(0， 1)，于是应用线性运算的性质，我们得到：

$$L(x,y)=L(x(1,0)+y(0,1))=xL(1,0)+yL(0,1)$$
$$=x(a,\ b)+y(c,\ d)=(ax,\ bx)+(cy,\ dy)$$
$$=(ax+cy,\ bx+dy).$$

这些信息可以总结在所谓的**矩阵方程**(matrix equation)中：

$$(x,\ y)\begin{pmatrix} a & b \\ c & d \end{pmatrix}=(ax+cy,\ bx+dy).$$

这里我们举了矩阵相乘的一个例子，它展示了这样的运算在一般情况下如何进行。**矩阵**(matrix)是由数组成的矩形阵列，它代表了另一种二维的数的对象。矩阵渗透了高等数学的几乎所有领域，同时包括纯数学和应用数学。它们代表了代数学中的一个重要部分。矩阵已经被证明有用到什么程度呢？现代数学的很大一部分便是努力要将自己用矩阵的语言表达出来。两个有相同行数和列数的矩阵可以逐项相加。比如，要找到两个矩阵之和的第二行第三列的元素，我们只需要将两个矩阵中相应位置的元素相加。不过，乘法赋予了矩阵一个新的重要特点，前面这个例子已经体现出了矩阵乘法的规则——积矩阵中的每个元素都是第一个矩阵的某行与第二个矩阵中的某列进行**点乘**(dot product)得到的。这是说，第一个矩阵的某行元素与第二个矩阵的某列元素分别相乘后再相加。

矩阵遵循代数里通常的法则，除了乘法交换律(commutativity of multiplication)。也就是说对于两个矩阵A和B，一般来说AB=BA是不

对的。然而，矩阵乘法是**可结合的**(associative)，这意味着书写任意长度的矩阵连乘时，不需要括号也不会引起歧义。

平面上的线性变换(linear transformation)通常是指绕着原点的旋转、相对于通过原点的直线的反射、相对原点的扩大或收缩，以及所谓的**剪切**(shear，或者说搓，slanting)。剪切是指让点平行于一根固定的轴移动，移动的长度与每个点到固定轴的距离成正比，就类似于一本书的各页可以依次滑动那样。对于任意一连串这样的变换，其效果都可以通过将所有对应的矩阵乘在一起来体现，这会给我们一个单独的矩阵，它包含了所有这些变换依次执行之后的最终效果。就像我们已经看到的那样，得到矩阵的各行正是两个点(1， 0)和(0， 1)经过变换后得到的像。这两个点称为**基向量**(basis vector)。

现在，对于代表绕原点逆时针旋转直角的矩阵J，我们自然地认为它应该模仿乘以虚单位i时我们观察到的数的行为。因为点(1， 0)被转到了点(0， 1)，类似地点(0， 1)移动到了(-1， 0)，所以这两个新向量组成了矩阵J的行。将J平方可得一个矩阵，它的几何效果是把点绕原点转过2× 90° = 180° 。下面我们就通过乘法来计算这个矩阵。例如，要找出J^2右下角的元素，我们取第二行和第二列的点乘，即(-1)× 1+0× 0 = -1+0 = -1。完整的计算如下：

$$J^2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I.$$

矩阵I的行是(1 0)和(0 1)，它是**单位矩阵**(identity matrix)。这么称呼是因为它就像数1一样，当乘另一个矩阵A时结果还是A。矩阵-I代表了绕原点半周，它的行为类似于-1，同时$(-I)^2$ = I。所有这些性质的最终结果是，对于实数a和b，矩阵aI+bJ在加法和乘法的意义

上很好地模仿了复数$a+bi$的行为，因此它给出了复数域的一个矩阵表示。对应于典型的复数$a+bi$的矩阵是

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix}.$$

代表复数的矩阵是满足乘法交换律的。当然就像之前提到的，这并不能推广到所有矩阵的乘积。另一个矩阵会出问题的地方是并非所有矩阵都能“逆转”。对于大部分方阵A(一个行数和列数相等的矩阵)，我们可能找到一个唯一的**逆矩阵**(inverse matrix)B，使得AB=BA=I。但是，逆矩阵存在与否，取决于一个单独的数。这个数与相应的方阵联系在一起，叫作它的**行列式**(determinant)。笼统地说，从矩阵的每行各取一个数，使它们占据不同的列，将它们相乘并赋予不同的符号，再将所有可能的这样的乘积相加，就得到了行列式。对于前文介绍的典型的2×2矩阵，行列式即为数$\Delta=ad-bc$。行列式用处很多，它具有优良的性质。例如：Δ代表了对应的变换的面积缩放因数：一个面积为a的形状经过一个行列式为Δ的矩阵所对应的变换，会变为面积为Δa的形状(倘若Δ为负，这个形状还会经过一次反射，将原始的朝向颠倒过来)。另外，两个方阵乘积的行列式，是这些矩阵行列式的乘积。在$\Delta=0$时，方阵A没有逆矩阵；除此情况之外，A有逆矩阵B。行列式为0在几何上对应于一个**退化的**(degenerate)变换，这时变换后得到的图形面积为0，比如说一条线段，甚至是一个单独的点。

对于一个复数$z=a+bi$的矩阵，我们注意到$\Delta=a^2+b^2$，它永远不为0，除了$z=0$的时候。当然，以前0从来没有倒数，这在更广阔的复数的世界里也仍然适用。不过，这也确认了每个**非零**的复数都有一个倒数。

到这里，我们已经站在了一个广阔世界的边缘，前方是线性代数、表示论[1]以及它们在多元微积分里的应用。我们就不再往远处走了。但是，读者应该知道矩阵其实不仅仅适用于三维，还适用于n维空间，通常这是通过n× n大小的矩阵实现的。虽然这些阵列变得更大更复杂了，但是矩阵自己依然是一个二维的数值对象。

复平面以外的数

在两个重要的意义上，所有复数组成的域C是**完备的**(complete)。假设有一个复数的无穷数列，它的各项往越来越小的圆圈里聚集，圆圈的半径趋向于0，我们称这个数列是**收敛的**(convergent)。任何复数的收敛数列都趋近于一个极限复数。这对实数来说也是真的，但不适用于有理数——对于任何无理数，其相继的小数近似都代表了一个有理数的数列，这个数列趋近于一个有理数范围之外的极限。另外，C在代数意义上是完备的(或者说是**封闭的**)，意思是可以证明，任何多项式方程$p(z)=a+bz+cz^2+\cdots+z^n=0$都有n个(复数)解：$z_1$， z_2，…， z_n，这使得$p(z)$自己可以彻底地因式分解为$p(z)=(z-z_1)(z-z_2)\cdots(z-z_n)$。

复数在这里和其他问题上获得了出人意料的成功，这在很大程度上消除了将数系进一步拓展到复平面以外的需求。事实上，要想构造一个范围更大的数字系统，从而在包含C同时又保留代数中所有的常规法则，这是不可能实现的。并且，只有两个推广的系统能做到保留一些代数结构，即**四元数**(quaternion)和**八元数**(octonion)。虽然它们的使用不如复数来得广泛，但是四元数在某些领域中得到了应用，比如三维计算机图形学。八元数可以看作一对四元数，它们不仅缺乏乘法的可交换性，甚至连乘法的可结合性质也丢失了。

一个四元数是形如$z=a+bi+cj+dk$的数，第一部分$a+bi$是一个普通的复数，同时两个**四元数单位**(quaternion unit) j和k满足$j^2 = k^2 = -1$。为了进行四元数的乘法，我们需要知道虚单位量如何相乘，这由以下规则决定：$ij = k$，$jk = i$，$ki = j$，不过反过来的积拥有相反的符号，比如$ji = -k$。其实，所有这些积都可以通过一个额外的方程：$ijk = -1$来推出。于是，四元数构成了一个增广了的代数系统，这个系统满足除乘法交换律以外的所有代数法则，失去交换律的原因在于上面提到的反向相乘时符号的变化。这个系统的相容性也可以通过2×2的矩阵表达来显示，不过这次我们不仅有实的元素，同时还允许出现复元素。数1再一次对应于单位矩阵I，但是单位量i，j和k则对应于矩阵：

$$\mathrm{i}=\begin{pmatrix}\mathrm{i} & 0\\ 0 & -\mathrm{i}\end{pmatrix},\ \ \mathrm{j}=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix},\ \ \mathrm{k}=\begin{pmatrix}0 & \mathrm{i}\\ \mathrm{i} & 0\end{pmatrix}.$$

而典型的四元数z有矩阵形式：

$$z=\begin{pmatrix}a+b\mathrm{i} & c+d\mathrm{i}\\ -c+d\mathrm{i} & a-b\mathrm{i}\end{pmatrix}.$$

然而，这种将四元数表为矩阵的方法不是唯一的。其实，复数的矩阵表达也有其他等价的形式。甚至，表示四元数还可以不用复数，只不过这要以使用更大的矩阵为代价：四元数可以表为某些只含有实数的4×4矩阵。

有时我们需要进行一些运算，其结果却不能被现有的数系所容纳，对这类计算的需求催生了新类型的数和对旧系统的推广。每个文明都是从自然数开始的，涉及数的片段的计算产生了分数，涉及债务的计算引出了负数，以及正如毕达哥拉斯发现的，涉及长度的计算产生了无理数。并非所有关于数的事务都能用整数以及它们的比值来处理，虽然这一发现已经十分久远，但这个事实依然深刻又微妙。随着科学变得越来越复杂，所需要的数的系统也必须成熟起来，才能处理这些进展。科学家们一般不会异想天开地主动创造新的数系，相反，在一开始，这些新数常常是被不情愿地、犹豫不决地引进来，用以应付科研中遇到的难题。例如，虽然在19世纪就被引入，矩阵直到20世纪初才在量子力学中取得了不容置疑的地位，当时的科学家们遇到了一个形如$q=AB-BA$却又不为0的量。在其他可交换的数系中，q当然会是0，因此这里需要的是一种他们以前从没遇到过的数值对象：它们是矩阵。

现在，看起来似乎数学和物理的世界已经拥有足够多种类的数了。虽然还存在本书没有提到的数的类型，但是自20世纪上半叶以来，我们还不需要对数学和科学中常用的数进行大的改造。

本书中我们对数学进行了一次热气球观光之旅，伴随着以上的观察结果，我们也到该说再见的时候了。我们从地面出发，渐渐升到了空中，我希望从这个高度上，丰富多彩又神秘莫测的数的世界能够吸引读者朋友的注目。

全书完

[1] 数学中抽象代数的一支。

术语对照

A

abundant number　盈数

Ahkmim papyrus　艾赫米姆莎草纸

AKS primality test　AKS素性测试

algebraic number　代数数

algorithm　算法

aliquot function　真因数和函数

aliquot sequence　真因数和数列

amicable pair　相亲数

Archimedes Cattle Problem　阿基米德牛群问题

Archimedes　阿基米德

argument (of complex number)　(复数的)辐角

arithmetic sequence (series)　算术数列

arithmetic triangle　算术三角形

associative　可结合的

asymmetric cipher　不对称密码

B

Babylon 巴比伦

base number 底数，基数

base ten 十进制，以十为底

base twelve 十二进制，以十二为底

basis vectors 基向量

Bell number 贝尔数

Bertrand’s Postulate 伯特兰-切比雪夫定理

Bertrand-Whitworth Ballot Problem 伯特兰-惠特沃斯选票问题

binary string 二进制字符串

binary 二进制

binomial coefficients 二项式系数

C

Cantor, Georg 格奥尔格·康托尔

Cantor’s diagonal argument 康托尔对角线论证法

Cantor’s middle third set 康托尔三分点集

Catalan number 卡特兰数

central binomial coefficients 中央二项式系数

cipher key 密钥

cipher 密码

code 代码

Collatz Algorithm 考拉兹算法

common difference 公差

common ratio 公比

complex plane 复平面

complex conjugate 复共轭

complex number 复数

complex 复

composite number 合数

compound interest 复利

continued fraction 连分数

convergents 收敛子

coprime 互素

countable set 可数集合

counting number　计数数

cryptography　密码学

cube　立方

D

decimals　小数

deciphering number　解密数

Dedekind cut　戴德金分割

deficient number　亏数

Delian Problems　提洛问题

denominator　分母

dense set　稠密集

determinant of a matrix　矩阵的行列式

determinant　行列式

Diophantine equation　丢番图方程

Dirichlet, Johann　约翰·狄利克雷

Distributive Law　分配律

divisor　除数，因子，因数

dot product (of vectors) (向量的)点乘

E

Egyptian fractions 埃及分数

enciphering number 加密数

Euclid of Alexandria 亚历山大里亚的欧几里得

Euclidean Algorithm 欧几里得算法

Euclidean number 欧几里得数

Euclidean Property 欧几里得性质

Eudoxus of Cnidus 尼多斯的欧多克索斯

Euler phi function 欧拉Φ 函数

Euler, Leonhard 莱昂哈德·欧拉

Euler’s Theorem 欧拉定理

exponential function 指数函数

e 自然常数e

F

factor 因数

factorials 阶乘

fast exponentiation　快速求幂

Fermat prime　费马素数

Fermat, Pierre de　皮埃尔·德·费马

Fermat’s Last Theorem　费马大定理

Ferrar’s graph　费勒斯图像

Fibonacci (of Pisa)　(比萨的)斐波那契

Fibonacci number　斐波那契数

fractions　分数

fundamental solution of pell equation　佩尔方程的基本解

G

Galilei, Galileo　伽利略·伽利雷

Gauss, Karl Friedrich　卡尔·弗里德里希·高斯

geometric series　几何数列

GIMPS　互联网梅森素数大搜索项目

Goldbach’s Conjecture　哥德巴赫猜想

Golden Ratio　黄金分割比

Golden Rectangle　黄金三角形

greatest common divisor 最大公约数

H

hailstone number 冰雹数

highest common factor 最大公因数

Hilbert Hotel 希尔伯特旅馆

Hindu-Arabic number system 印度-阿拉伯数字系统

hyperbolic function 双曲函数

I

identity matrix 单位矩阵

imaginary axis 虚轴

imaginary number 虚数

inclusion-exclusion principle 容斥原理

incommensurable 不可公度的

integers 整数

inverse matrix 逆矩阵

irrational number 无理数

L

lattice point　栅格点

linear equation　线性方程

linear operation　线性运算

logarithm　对数

M

matrix equation　矩阵方程

matrix multiplication　矩阵乘法

matrix　矩阵

measure of set　集合的测度

Mersenne number　梅森数

modular arithmetic　模算术

modulus of complex number　复数的模

multiplicative identity　乘法单位量

N

natural logarithm　自然对数

natural number　自然数

negative number　负数

neighbourhood　领域

nowhere dense　处处不稠密

number line　数轴

number　数

numerator　分子

O

octonians　八元数

one-time pad　一次性密码本

origin　原点

P

partition number　分拆数

Pascal’s triangle　帕斯卡三角形

Pell equation　佩尔方程

perfect number　完美数

pi　圆周率π

plaintext　明文

polar co-ordinates　极坐标

pole　极点

polynomial　多项式

positional value　位值进位法

positive number　正数

prime divisibility tests　素性测试

prime factorization　素因数分解

prime number　素数

prime　素数，素的

progression　数列

proper divisor　真因数

proper factor　真因数

public key　公共密钥

Pythagoras’s theorem　勾股定理

Pythagorean triple　勾股数

Q

quadratic equation　二次方程

quaternions　四元数

R

Rabbit Problem 兔子问题

radian 弧度

rational number 有理数

real axis 实轴

real number 实数

recurrence relation 递推关系

recurring decimal 循环小数

reduced fraction 既约分数

refiection principle 反射原理

remainder 余数

Riemann Hypothesis 黎曼猜想

Roman numerals 罗马数字

RSA RSA 加密过程

Russell’s Paradox 罗素悖论

S

self-similarity 自相似

semiperfect number　半完美数

set　集合

shear　剪切

Sieve of Eratosthenes　埃拉托斯特尼筛法

sociable number　多亲数

square number　平方数

square root of 2　2的平方根

square root of minus 1　-1的平方根

square root　平方根

square-free number　无平方数因数的数

Stirling formula for factorials　阶乘的斯特林公式

Stirling number　斯特林数

Stirling, James　詹姆斯•斯特林

straightedge and compass　直尺和圆规

subset　子集

symmetric cipher　对称密码

T

terminating decimal　有限小数

ternary　三进制

transcendental number　超越数

triangular number　三角数

trigonometric identities　三角恒等式

twin primes　孪生素数

uncountable set　不可数集合

unit fraction　单位分数

V

vector　向量

W

weird number　奇异数

Wilson’s Theorem　威尔逊定理

Y

Young diagram　杨表

Z

ZFC set theory　包含选择公理的策梅罗-弗兰克尔集合理论

Peter M. Higgins

NUMBERS

A Very Short Introduction

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

© Peter M. Higgins 2011

The moral rights of the author have been asserted
Database right Oxford University Press (maker)

First Published 2011

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organization. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above

You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Typeset by SPI Publisher Services, Pondicherry, India
Printed in Great Britain on acid-free paper by
Ashford Colour Press Ltd, Gosport, Hampshire

ISBN 978–0–19–958405–5

1 3 5 7 9 10 8 6 4 2

Contents

Preface

The purpose of this little book is to explain, in language that will be familiar to everyone, what are the various kinds of numbers that arise and how they behave. Numbers allow comparisons between all manner of things, and anyone with no understanding of numbers would be lost in the modern world where things numerical are there to greet us at every turn. We should realize however that, despite their familiarity, numbers have no physical existence but rather are abstractions that we elicit from the world wefind ourselves in. To develop a clear picture of how they operate, it is sometimes better to consider them in their own right,without reference to anything else.

This Short Introduction is not a refresher course in arithmetic, nor will that much be said on the history of the number system.Rather its purpose is to explain numbers themselves and the kinds of behaviour they exhibit. A glance at the list of chapters reveals that thefirst part of the book deals mainly with ordinary counting numbers, while in the second half we go beyond that. By exploring natural problems that arise in commerce and science, the need to freely perform calculations eventually has taken us, by stages, into the arena of the complex numbers, which is the principal underlying framework for most number matters. This may sound a little daunting but rest assured, the hard work has already been done for you.

The modern number system did not come to us gift-wrapped but rather has developed over many centuries. There were long periods of confusion, which had two root causes. Thefirst was the lack of an efficient way to represent the numbers we need that would allow us to manipulate them. The second, which was related to thefirst, was philosophical agonizing over the interpretations of various number types and whether or not they are meaningful. Nowadays we are much more sure of ourselves when it comes to what we do and don’t need to worry about when dealing with numbers, making it possible to give a complete picture of the number world in a single short account like this one.That is not to say that all mystery has vanished- far from it, as you will discover as you read on.

Peter M. Higgins

Colchester, England,2011

List of illustrations

1Primesieve 11

2 The Arithmetic Triangle 56

3 Catalan numbers and mountain patterns 59

4 Pentagon and the Golden Rectangle 62

5 Stirling’s Triangle 63

6 Dual partitions of 17 66

7 Central portion of the number line near 0 70

8 The evens and the squares paired with the natural numbers 86

9 Cantor’s Diagonal Argument 89

10 Rationals separate any given positions on the number line 93

11 Evolution of Cantor’s Middle Third Set to the 4th level 95

12 Lattice points on the line of a linear equation 99

13 Addition of complex numbers by adding directed line segments 111

14 The position of a complex number in polar coordinates 113

15 Multiplication by i rotates a complexnumberbyaright angle 117

Chapter 1 How not to think about numbers

We are all used to seeing numbers written down, and to drawing some meaning from them. However, a numeral such as 6 and the number that it represents are not one and the same thing. In Roman numerals, for example, we would write the number six as VI, but we realize that this stands for the same number that is written as 6 in modern notation. Both symbolize collections of the kindcorrespondingtosixtallymarks: IIIIII.Weshallfirst spend a little time considering the different ways that we represent and think about numbers.

We sometimes solve number problems almost without realizing it.For example, suppose you are conducting a meeting and you want to ensure that everyone there has a copy of the agenda. You can deal with this by labelling each copy of the handout in turn with the initials of each of those present. As long as you do not run out of copies before completing this process, you will know that you have a sufficient number to go around. You have then solved this problem without resorting to arithmetic and without explicit counting. There are numbers at work for us here all the same and they allow precise comparison of one collection with another, even though the members that make up the collections could have entirely different characters, as is the case here, where one set is a collection of people, while the other consists of pieces of paper.

What numbers allow us to do is to compare the relative size of one set with another.

In the previous scenario you need not bother to count how many people were present as you did not have to know- your problem was to determine whether or not the number of copies of the agenda was at least as great as the number of people, and the value of these numbers was not required. You will, however, need to take a count of the number present when you order lunch forfifteen and certainly, when it comes to totting up the bill for that meal, someone will make use of arithmetic to work out the exact cost, even if the sums are all done on a calculator.

Our modern number system allows us to express numbers in an efficient and uniform manner, which makes it easy to compare one number with another and to perform the arithmetical operations that arise through counting. In the day-to-day world, we employ base ten for all our arithmetic, that is to say we count by tens, and we do that for the accidental reason that we have ten digits on our hands. What makes our number system so effective, however, is not our particular choice of base but rather the use of positionalvalue in number representation, where the value of a numeral depends on its place in the number string. For example, 1984 is short for 4 ones plus 8 lots of ten plus 9 hundreds plus 1 thousand.

It is important to understand what we mean when we write numbers in particular ways. In this chapter, we will think about what numbers represent, discover different approaches to counting, meet a very important set of numbers(the primes), and introduce some simple number tricks forfinding them.

Howcountingwassortedout

It is worth taking a few moments to appreciate that there are two distinct stages to the process of building a counting system based on, for instance, tens. Two basic tasks that we impose on children are remembering how to recite the alphabet and learning how to count. These processes are superficially similar but yet have fundamental differences. Our language is based on a 26-letter alphabet and, roughly speaking, each letter corresponds to a sound that we use to speak words. In any case, it is certainly true that the English language has developed so that it can be written using a set of 26 symbols. However, we cannot compile dictionaries unless we assign an order to our alphabet. There is no particularly natural order available and the one that we have settled on and all sing in school, a, b, c, d, ... seems very arbitrary indeed. To be sure, the more frequently used letters generally occur in thefirst half of the alphabet, but this is only a rough guide rather than a rule, with the common letters s and t, forexample, sounding off late in the roll call. By contrast, the counting numbers, or natural numbers as they are called, 1, 2, 3, ... come to us in that order: for

example, the symbol 3 is meant to stand for the number that follows 2 and so has to be listed as its successor. We can, up to a point, make up a fresh name for each successive number. Sooner or later, however, we have to give up and start grouping numbers in batches in order to handle the unending sequence. Grouping by tens represented thefirst stage of developing a sound number system, and this approach has been near universal throughout history and across the globe.

There was, however, much variation in detail. The Roman system favours gathering byfives as much as tens, with special symbols, V and L, forfive and forfifty respectively. The Ancient Greek system was squarely based on grouping by tens. They would use specific letters to stand for numbers, sometimes dashed to tell the reader that the symbol should be read as a number rather than as a letter in some ordinary word. For example, π stood for 80 andγ for 3, so they might writeγ π to denote 83. This may look equally as efficient, and indeed much the same as our notation, but it is not. The Greeks still missed the point of the positional system as the value of each of their symbols wasfixed. In particular, γ π could still only represent the same number, 3+80, whereas if we switch the order of the digits in 83 we have the different number 38.

In the Hindu-Arabic system, the second stage of number representation was attained. Here the big idea is to make the value of a symbol dependent upon where it occurs in the string. This allows us to express any number with just afixed

family of symbols. We have settled on the set of ten numerals 0, 1, 2, • • •, 9, so the normal number system is described as base ten, but we could build our number system up from a larger or a smaller collection of basic symbols. We can even manage with as few as two numerals, 0 and 1 say, which is what is known as the binary system, so often used in computing. It is not the choice of base size, however, that was revolutionary but the idea of using position to convey extra information about the identity of your numbers.

For example, when we write a number like 1905, the value of each digit depends on its place in the number string. Here there are5 units, 9 lots of one hundred(which is 10× 10), and one lot of one thousand(which is 10× 10× 10). The use of the zero symbol is important as a placeholder. In the case of 1905, no contribution comes from the 10's place, but we cannot ignore that and just write195 instead, as that represents an entirely different number. Indeed, each string of digits represents a different number and it is for that reason that huge numbers may be represented by short strings. For instance, we can assign a unique number to every human being on Earth using strings no longer than ten digits and in this way give a personal identifier to every individual belonging to this huge set.

Societies of the past sometimes used different bases for their writing of numbers but that is much less significant than the fact that nearly all of them lacked a true positional system with full use of a zero symbol as a

placeholder. In view of how very ancient is the civilization of Babylon, it is remarkable that they among the peoples of the ancient world came closest to a positional system.

They did not, however, fully embrace the use of the not-so-natural number 0 and eschewed using the empty register in thefinal position the way we do to distinguish, for example,830 from 83.

The conceptual hurdle that had to be cleared was the realization that zero was indeed a number. Admittedly, zero is not a positive number but it is a number all the same and until we incorporate it into our number system in a fully consistent manner, we remain handicapped. This crucialfinal step was taken in India in about the 6th century AD. Our number system is called Hindu-Arabic as it was communicated from India to Europe via Arabia.

Living with and without decimals

Of course, we have now extended the base ten positional idea into fractional parts giving the familiar decimal number system. When we write 3.14, for instance, the 1 after the decimal point stands for one lot of $\frac{1}{10}$, a tenth part, and the 4 similarly stands for $\frac{4}{100}$. This two-decimal-place form of a number is very familiar to us as we deal in decimal currency where the smallest unit is not one dollar, or pound, or euro, as the case may be, but the 'penny' or 'cent', which is one-hundreth part of the main currency unit. Decimal arithmetic is the natural extension of the base ten system and in

practice it represents the best way to carry out ordinary sums. Despite all its advantages, the decimal approach had a slow and hesitant genesis. It remained within the province of a mathematical elite until the latter part of the 16th century, when it finally found its way into commercial arithmetic and general usage. Even after that time, grouping based on numbers other than powers of ten persisted. Britain did not adopt decimal currency until 1971. And part of the English-speaking world still resolutely sticks with yards, feet, and inches. In defence of Imperial Units, they are convenient in size, being very much attuned to the human scale of things. Our hands are six to eight inches long and we arefive to six feet tall, so we surround ourselves with objects of similar size, which are then readily measured in units of feet and inches. However, ten inches to the foot could have worked just as well and would have been more easily dealt with by our base ten calculators.

Adopting a particular base for a number system is a little like placing a particular grid scale on a map. It is not intrinsic to the object but is rather akin to a system of coordinates imposed on top as an instrument of control. Our choice of base is arbitrary in character and the exclusive use of base ten saddles us all with a blinkered view of the set of counting numbers: 1, 2, 3, 4, • • •. Only by lifting the veil can we see numbers face to face for what they truly are. When we mention a particular number, let us say for example forty-nine, all of us have a mental picture of the two numerals 49. This is somewhat unfair to the number in

question as we are immediately typecasting forty-nine as(4× 10)+9. Since49=(4× 12)+1, it may just as easily be thought of that way and, indeed, in base twelve, forty-nine would therefore be written as 41, with the numeral 4 now standing for 4 lots of 12. However, what gives the number forty-nine its character is that it equals the product 7× 7, known as the square of 7. This facet of its personality is highlighted in base seven, as then the number forty-nine is represented as 100, the 1 now standing for one lot of7× 7. We would be equally entitled to use another base, such as twelve, for our number system: the Mayans used twenty and the Babylonians base sixty. In one way, the number 60 is a good choice for a counting base as 60 has many divisors, being the smallest number divisible by all the numbers from 1 through to 6. A relatively large number such as 60 has the disadvantage, however, that to use it as a base would require us to introduce 60separate symbols to stand for each of the numbers from zero up to fifty-nine.

One number is a factor of another if thefirst number divides into the second a whole number of times. For example, 6 is a factor of42=6× 7 but 8 is not a factor of 28 as 8 into 28 goes 3 times with a remainder of 4. The property of having many factors is a handy one to have for the base of your number system, which is why twelve may have been a better choice than ten for our number base as 12 has 1, 2, 3, 4, 6, and 12 as its list of factors while 10 is divisible only by 1, 2, 5, and 10.

The effectiveness and sheer familiarity of our number system embues us with a false confidence and with some inhibitions. We feel happier with a single whole number than with an arithmetical expression. For example, most people would rather talk about5969 than 47× 127, although the two expressions represent the same thing. Only after ‘working out the answer’, 5969, do we feel that we ‘have’ the number and can look it in the eye. There is, however, an element of delusion in this as we have only written the number as a sum of powers of ten. The general shape of the number and other properties can be inferred more from the alternative form where the number is broken down into a product of factors. To be sure, this standard form, 5969, does allow direct comparison with other numbers that are expressed in the same way but it does not reveal the full nature of the number. In Chapter 4, you will see one reason why the factorized form of a number can be much more precious than its base ten representation, which can keep vital factors hidden.

One advantage that the ancients did have over us is that they were not mentally trapped within a decimal-style mindset. When it came to number patterns, it was natural for them to think in terms of special geometric properties that a particular number may or may not enjoy. For example, numbers such as 10 and 15 aretriangular, something that is visible to us through the triangle of pins in ten-pin bowling and the triangular rack offifteen red balls in snooker. But this is not something that comes to mind from the base ten displays of these numbers alone. The freedom the ancients enjoyed by

default we can recapture by casting aside our base ten prejudices and telling ourselves that we are free to think of numbers in quite different ways.

Having emancipated ourselves in this way, we might choose to focus on factorizations of a number, that is to say the way the number can be written as a product of smaller numbers multiplied together. Factorizations reveal something of the number's inner structure. If we suspend the habit of thinking of numbers simply as servants of science and commerce, and take a little time to study them in their own right without reference to anything else, much is revealed that otherwise would remain hidden. The natures of individual numbers can manifest themselves in ordered patterns in nature, more subtle than mere triangles and squares, like the spiral head of a sunflower, which represents a so-called Fibonacci number, a number type that will be introduced in Chapter 5.

Aglance attheprimenumbersequence

One of the glories of numbers is so self-evident that it may easily be overlooked- every one of them is unique. Each number has its own structure, its own character if you like, and the personality of individual numbers is important because when a particular number arises, its nature has consequences for the structure of the collection to which that number applies. There are also relationships between numbers that reveal themselves when we carry out the

fundamental number operations of addition and multiplication. Clearly any counting number greater than 1 can be expressed as the sum of smaller numbers. However, when we start multiplying numbers together, we soon notice that there are some numbers that never turn up as the answer to our sums. These numbers are the primes and they represent the building blocks of multiplication.

A prime number is a number like 7 or 23 or 103, which has exactly two factors, those necessarily being 1 and the number itself.(The word divisor is also used as an alternative word for factor.)We do not count 1 as a prime as it has only one factor. Thefirst prime then is 2, which is the only even prime and the following trio of odd numbers 3,5, and 7 are all prime. Numbers greater than 1that are not prime are called composite as they are composed of smaller numbers. The number $4=2\times2=2^2$ is thefirst composite number;9 is thefirst odd composite number, and $9=3^2$ is also a square. With the number $6=2\times3$, we have thefirst truly composite number in that it is composed of two different factors that are greater than 1 but smaller than the number itself, while$8=2^3$ is thefirst proper cube, which is the word that means that the number is equal to some number raised to the power 3.

After the single-digit numbers, we have our chosen number base$10=2\times5$, which is special nonetheless being triangular in that$10=1+2+3+4$(remember ten-pin bowling). We then have a pair of twin primes in 11 and 13, which are two consecutive odd numbers that are both prime, separated by the number 12,

which in contrast has many factors for its size. Indeed, 12 is thefirst so-called abundant number, as the the sum of its proper factors, those less than the number itself, exceeds the number in question: 1+2+3+4+6=16. Thenumber14=2× 7maylook undistinguished but, as the paradoxical quip goes, being the firstundistinguished number makes it distinguished after all. In15=3× 5, we have another triangular number and it is thefirst odd number that is the product of two proper factors. Of course, $16=2^4$ is not only a square but thefirst fourth power(after 1), making it very special indeed. The pair 17 and 19 are another pair of twin primes, and I leave the reader to make their own observations about the peculiar nature of the numbers 18, 20, and so on. For each you can make a claim to fame.

Returning to the primes, thefirst twenty of them are:

2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71.

Clearly, near the very beginning of the number sequence, primes are commonplace as there is little opportunity for small numbers to have factorizations. After that, the primes become rarer. For example, there is only one triple of consecutive odd primes: the trio 3, 5, 7 is unique, as every third odd number is a multiple of 3, and so this can never happen again. The thinning process of prime occurrence is, however, quite leisurely and surprisingly erratic. For example, the thirties have only two primes, those being 31 and 37, yet immediately after 100 there are

two ‘consecutive’ pairs of twin primes in 101,103 and 107,109.

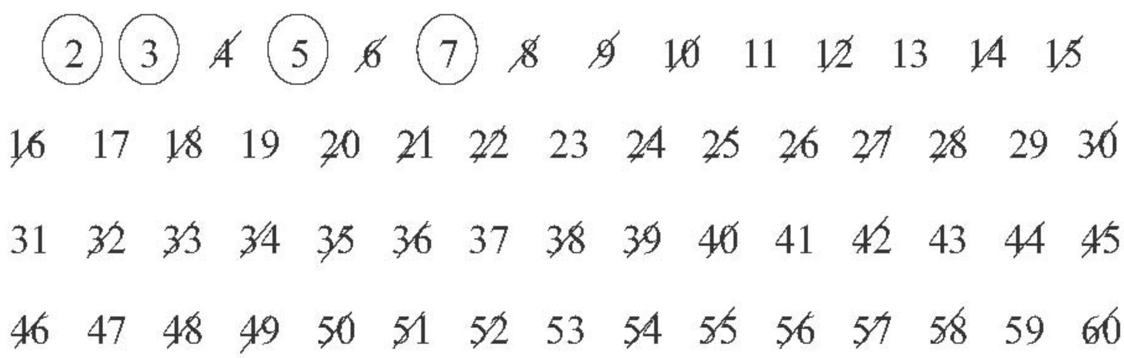
The primes have been a source of fascination for thousands of years because they never run out(a claim that we shall justify in the next chapter)yet they arise among the natural numbers in a somewhat haphazard fashion. This mysterious and unpredictable facet of their nature is exploited in modern cryptography to safeguard confidential communication on the Internet, which is the subject of Chapter 4.

Checking for primality:prime divisibility tests

The most simple-minded way offinding all the primes up to a given number such as 100 is to write all the numbers down and cross off the composite numbers as youfind them. The standard method based on this idea is called the Sieve of Eratosthenes and runs as follows. Begin by circling 2 and then cross off all the multiples of 2(the other even numbers)in your list. Then return to the beginning, circle thefirst number you meet that has not been crossed off(which will be 3)and then cross off all its multiples in the remaining list. By repeating this process sufficiently often, the primes will emerge as those numbers that never get crossed out, although some will be circled and some not.For example, Figure 1 shows the workings of the sieve up to 60.

How do you know when you can stop sieving?You need to repeat this process until you circle a number that is greater

than the square root of the largest number in your list. For instance, if you do your own sieve for all numbers up to 120, you will need to run through the sieve for multiples of 2,3,5, and 7, and when you circle 11 you can stop, as 112=121. At that point, you will have circled as far as thefirst prime exceeding the square root of your biggest number(120 in this case)with the remaining primes sitting there untouched. All composite numbers will now have been crossed out as each is a multiple of one or more of 2,3,5,and 7.



1. Prime sieve: the primes up to 60 are the numbers notcrossed out

It is not hard to see why the square root of the greatest number n in your list determines the number of passes you need to make.(When explaining properties of arbitrary numbers,mathematicians give names, in the form of symbols, for the subjects of the discussion. For numbers, these names are typically lower-case letters, such as m and n; the product of two numbers $m \times n$ is often abbreviated as mn.)Any listed composite number m will have a prime factor and its smallest

prime factor must be no more than the square root of n, because the product of two or more numbers that exceed $\sqrt{n}$ is greater than n (and so also greater than m).

Another aspect of the question of primality is whether a particular given number n is prime or composite. To decide that, we can test n for division by each of the primes in turn up to $\sqrt{n}$ and if n passes all these tests it will be prime, and otherwise not. For that reason, it is handy to have some simple ways of testing for divisibility by each of the small primes, 2, 3, 5, 7, ... This need is easily catered for as follows.

It is very easy to test for divisibility by 2 and by 5 as these primes are the prime factors of our number base ten. In view of this, you only need to check thefinal digit of the number n in question: n is divisible by 2 exactly when its units digit is even (i.e. 0, 2, 4, 6, or 8), and n has 5 as a factor if and only if it ends in 0 or 5. No matter how many digits the number n has, we only need to check the last digit to determine whether we have a multiple of 2 or of 5. For primes that do not divide into 10, we need to do a bit more work but nevertheless there are simple tests for divisibility that are much quicker than resorting to doing the full division sum.

A number is divisible by 3 if and only if the same is true of the sum of its digits. For example, the sum of the digits of $n = 145,373,270,099,876,790$ is 87 and $87 = 3 \times 29$ and so n is in this case divisible by 3. Of course, we can apply the

test to the number 87 itself and indeed go on taking the sum of digits of the outcome at each stage until the result is obvious. Doing this for the given example produces the following sequence:

$$145,373,270,099,876,790 \rightarrow 87 \rightarrow 15 \rightarrow 6=2\times 3.$$

You will see that all the division tests listed here are so quick that you can handle numbers with dozens of digits with relative ease even though these numbers are billions of times greater than the biggest number with which your hand calculator can cope.

The tests given here for the remaining primes up to 20 are chosen because they are all of the same general type. These routines are all simple to apply, although it is less obvious why they work.Although the justifications are not recorded here, the proofs of their validity are not especially difficult.

Let’s begin with a test for divisibility of a given number n by 7.Double thefinal digit of n and subtract that from the number that remains when thatfinal digit is removed. The new number will be a multiple of 7 precisely when the same is true of n. We repeat this process until the outcome is obvious. As a simple example, let us take n=3465: twice 5 is 10 so we take 10 from the 346 to get 336; next we go again, taking twice 6, which is 12 from the 33 to get21=3× 7, and so nis divisible by 7. If you have forgotten your seven times tables, we can go through again: subtracting twice 1from 2

leaves us with 0, which is divisible by 7, as $\frac{0}{7}=0$. (Itisfine to divide zero by a whole number- it is the reverse, dividing by zero, that has no meaning.)Even a number in the tens of millions can be dealt with easily in this way. In this and subsequent examples, we simply list the output number at each stage of thealgorithm, which is the name given for a mechanical process such as this one that solves a given type of problem.

n=27, 916, 924→2, 791, 684→279, 160→27, 916→2, 779→259→7

and so n is divisible by 7. Each time we run through the loop of instructions, we lose at least one digit, so the number of passes through the loop is about the same as the length of the number with which we begin.

To test whether or not n has a factor of 11, subtract thefinal digit from the remaining truncated number and repeat. For example, the next number is a multiple of 11 as our method reveals:

4, 959, 746→495, 968→49, 588→4, 950→495→44=4× 11.

To check for divisibility by 13, add four times thefinal digit to the remaining truncated number and precede as with 7 and 11. For instance, the next number turns out to have 13 as one of its prime factors:

11, 264, 331→1, 126, 437→112, 671→11, 271

→1131→117→39=3× 13. For 17 and for 19, we subtractfive times thefinal digit in the case of 17, and add twice thefinal digit when testing if 19 is a factor, once more applying this step to the truncated number that remains, repeating the process as often as we need. For example, we test 18,905 for divisibility by 17:

18,905→1,865→161→11

so it is not a multiple of 17, but for 19, the test gives the opposite conclusion:

18,905→1,900=100× 19.

Armed with this battery of tests, you can readily check the primality of all numbers up to 500(as 23^2=529 exceeds 500, so19 is the largest potential prime factor that you need concern yourself with). For example, to settle the matter for 247, we just need to check for divisibility up to the prime 13(as the square of the next prime, 17^2=289, exceeds 247). Applying the test for 13, however, we learn from 247→(24+28)=52→13, that we have a multiple of 13: (247=19× 13).

The divisibility tests for primes can also be mounted in parallel to furnish divisibility tests for those numbers that are square-freeproducts of these primes(numbers not divisible by the square of any prime)such as 42=2× 3× 7: a number nwill be divisible by42 exactly when npasses the trio of divisibility tests for 2, 3, and 7. However, tests for those

numbers that have square factors, such as $9=3^2$, do not come automatically, although it is the case that n has 9 as a factor if and only if that is true of the sum of the digits of n.

You might ask, after thousands of years, haven't those clever mathematicians come up with better and more sophisticated methods of testing for primality?The answer is yes. In 2002, a relatively quick way was discovered to test if a given number is prime. The so-called 'AKS primality test' does not, however,provide the factorization of the given number if it happens to be composite. The problem offinding the prime factors of a given number, although in principle solvable by trial, still seems practically intractable for extremely large integers, and for that reason it forms the basis of much ordinary encryption on the Internet, a subject to which we will return in Chapter 4. Before that we shall, in the next two chapters, look a little more closely at primes and factorization.

Chapter 2 The unending sequence of primes

How primesfit into the number jigsaw

How can we be sure that the primes do not become rarer and rarer and eventually peter out altogether?You might think that since there are infinitely many counting numbers and each can be broken down into a product of primes(something explained more carefully in a moment), there must then be infinitely many primes to do the job. Although this conclusion is true, it does not follow from the previous observations, for if we begin with afinite collection of primes, there is no end to the number of different numbers we can produce just using those given prime factors.Indeed, there are infinitely many different powers of any single prime:for example, the powers of the prime 2 are2,4,8,16,32,64,・・・. It is conceivable therefore that there are onlyfinitely many primes and every number is a product of powers of those primes. What is more, we have no way of producing an unending series of different primes the way we can,for example, produce any number of squares, or multiples of a specific given number. When it comes to primes, we still have to go out hunting for them, so how can we be sure they do not become extinct?

We will all be sure by the end of this chapter, butfirst I will draw your attention to one simple ‘pattern’ among the primes worth noting. Every prime number, apart from 2 and 3,

lies one side or the other of a multiple of 6. In other words, any prime after these firsttwohastheform$6n\pm 1$ for some number n.(Remember that$6n$is short for $6\times n$ and the double symbol$\pm$means plus or minus.)The reason for this is readily explained. Every number can be written in exactly one of the six forms $6n$, $6n\pm 1$, $6n\pm 2$, or$6n+3$ as no number is more than three places away from some multiple of six. For example, $17=(6\times 3)-1$, $28=(6\times 5)-2$, $57=(6\times 9)+3$; indeed, the six given forms appear in cyclic order, meaning that if you write down any six consecutive numbers, each of the forms will appear exactly once, after which they will reappear again and again, in the same order. It is evident that numbers of the forms $6n$and $6n\pm 2$ are even, while any number of the form $6n+3$ is divisible by 3. Therefore, with the obvious exceptions of 2 and 3, only numbers of the form $6n\pm 1$canbe prime. Thecasewhereboth of the numbers $6n\pm 1$areprime corresponds exactly to the twin primes: for example$(6\times 18)\pm 1$gives the pair 107, 109 mentioned in thefirst chapter. You might be tempted to conjecture that at least one of the two numbers$6n\pm 1$ is always prime- this is certainly true for the list of primes up to 100 but thefirst failure is not far away: $(6\times 20)-1=119=7\times 19$, while$(6\times 20)+1=121=112$, so neither number is primewhenwetake$n=20$.

And the principal reason why primes are important is that every number can be written as a product of prime numbers, and that can be done in essentially only one way. Tofind this special factorization, we just need to factorize the given number in some way and then continue factorizing any

composite factors that appear until this can be done no more. For example, we could say that120=2× 60 and continue by breaking the composite factor of 60 down further to give:

120=2× 60=2× (2× 30)=2× 2× (2× 15)=2× 2× 2× 3× 5.

We say that the prime factorization of120is23× 3× 5. We could,however, have came to this by another route. For instance but rearranging the prime factors from least to greatest still yields the same result as before: 120=2^3× 3× 5.

120=12× 10=(3× 4)× (2× 5)=(3× (2× 2))× (2× 5)

At least it did in that example, and this behaviour may be more or less familiar to you, but how can you be sure that this applies to every number?It is clear enough that any number can be broken down into a product of primes but, since there is in general more than one way of tackling this task, how can we be sure that the process will always deliver the samefinal result?This is an important question, so I will take a few moments to give an outline of the reasoning that allows us to be absolutely sure about this. It is a consequence of another special property of prime numbers that we shall call the euclidean property: ifaprime number is a factor of a product of two or more numbers, then it is a factor of one of the numbers in that product. For example,7 is a factor of 8× 35=280(as the product 280=7× 40)and we note that 7 is a factor of 35. This property characterizes primes as no composite number can give you the same guarantee: for

example, we see that 6 is a factor of 8× 15=120(as 120=6× 20) yet 6 is not a factor of either 8 or 15.

The fact that primes always have the above property can be proved using an argument based on what is known as the Euclidean Algorithm, which will be explained in Chapter 4. If we take this on trust for the time being, it is not too difficult to explain why no number could have two different prime factorizations, for suppose there were such a number. There then would be a smallest one that behaved in this way: let us denote it by n and so nhas two prime factorizations which, when the prime factors are written in ascending order, are not identical. We shall show that this leads to contradiction and so must be false.

If these two factorizations of n had a prime p in common, we could cancel p from both and obtain two different prime factorizations of the smaller number np. Sincenis the smallest number with two distinct prime factorizations, this is not possible and so the sets of primes involved in each factorization of nmust have no prime in common. Now take one prime p that occurs in one of these factorizations of n. Since this prime p is a factor of n, it follows that p is a factor of the second factorization, and so, by the euclidean property, p is a factor of one of the primes, q say, of the second factorization. But since q and p are both prime, thisis only possible if q=p, a possibility that we have already discounted as the two factorizations of n have no common prime. And so we arrive at afinal contradiction, showing that

it is impossible for such a number nto exist. Therefore we conclude that the prime factorization of every number is unique.

It is worth noting that the uniqueness of prime factorization would not hold if we included the number 1 among the primes, as we can adjoin any power of 1 to a factorization and the product retains the same value. This shows that 1 is fundamentally different in nature to the primes, and so it is right to frame the definition of prime number in a way that excludes 1 from the collection.

Euclid’s infinity of primes

Let us return to the question as to how we know that the primes go on forever and that there is no way past them. If someone claimed that 101 is the largest prime, you can refute him at once by showing that 103 has no factors(except for 1 and 103)and so 103is a larger prime. Your friend might then concede that he made a slip and that he should have said that it was 103 that is the largest prime of all. You could then show him up again by demonstrating that 107 is also prime, but your friend might still persist in his error by adjusting his position to the largest prime number on view. He could even retreat a little further and admit that he does not know the identity of the largest prime but nevertheless continue to claim that he is certain that there is one.

The best way to settle this question would be to show that, given any conceivablefinite collection of primes, we

can produce a new prime not in the list. For example, if someone claimed there was a largest odd number out there somewhere, you could refute him by saying that if nis odd, then n+2 is a larger odd number, so there cannot be a largest odd number. This approach, however, is not so easy for the primes- given afinite list of primes, we have no way of using the collection to manufacture a prime that is demonstrably bigger than all of them. Perhaps there is a biggest prime after all?How are we to know that our stubborn friend isn’t right?

Euclid of Alexandria(c.300 BC), the Greek mathematician and father of all things euclidean, did however know. Given a list p_1, p_2, • • •, p_k where each of the pi denotes a different prime, he could notfind a way of generating a new prime, so he reverted to an argument that is one step more subtle. He showed that there must be one or more new primes within a certain range ofnumbers(but his argument does not allow us to locate exactly where tofind a prime within that range).

It goes like this. Let p_1, p_2, • • •, p_k be the list of thefirst kprimes say, and consider the number nthat is one more than the product of all these primes, so that $n=p_1p_2 \cdots p_k+1$. Eithernis a prime, or is divisible by a prime smaller than itself, which cannot be any of p_1, p_2, • • •, p_k, asifp is any one of these primes, then dividing nby p will leave a remainder of 1. It follows that any prime divisor of nis a new prime that is greater that all the primes p_1, p_2, • • •, $p_k$and no more than nitself. In particular, it

follows from this that there can be nofinite list of primes that contains every prime number, and so the sequence of primes continues on forever and will never be exhausted. Euclid’s eternal proof of the infinity of primes is among the most admired in all of mathematics.

Although Euclid’s argument does not tell exactly where tofind the next prime number, the overall frequency of the primes is now quite well understood. For example, if we take any two numbers, a and b say, with no common factor and consider the sequence a, a+b, a+2b, a+3b, •••, it was shown by the German mathematician Johann Dirichlet(1805- 59)that infinitely many members of such a sequence are prime.(Of course, there is no hope if a and b do have a common factor, d say, as then every member in the list is also a multiple of d, and so is not prime.)When a=1andb=2, we get the sequence of odd numbers which we know, by Euclid’s proof, contains infinitely many prime numbers. Indeed, it can be shown through fairly simple adaptations of Euclid’s argument that other special cases such as the sequence of numbers of the forms 3+4n, 5+6n, and5+8n(as nruns through the successive values 1, 2, 3, •••), each have infinitely many primes. The general result of Dirichlet is, however, very difficult to prove.

Another simply stated result is that there is always at least one prime number greater than any given number nbut less than 2n(for n≥2).(As an aid to memory, inequality signs such as this one, which stands for greater than or equal to,

always point to the smaller quantity.)This fact, historically known as Bertrand’s Postulate, can be proved using quite elementary mathematics,although the proof is itself quite tricky. We can verify the postulate for n up to 4000 by making use of the following list of primes.First observe that each number in the list after the initial prime 2is smaller than twice its predecessor:

2, 3, 5, 7, 13, 23, 43, 83, 163, 317, 631, 1259, 2503, 4001.

For each n in the range up to 4000, take the largest prime p in the list that is no more than n; the next prime q then lies in the range $n < q < 2n$ and this then ensures that Bertrand’s Postulate holds for all n up to 4000. For example, for n=100, p=83, and then $q=163 < 2\times 100$. A subtle argument involving the size of the so-called central binomial coefficients(introduced in Chapter 5)then shows that the postulate is also true for n larger than 4000.

However, we do not have to go too far before meeting similar-sounding problems that as yet remain unsolved. For example, no one knows if there is always a prime between any two consecutive squares. Another observation is that there seems to be enough primes to ensure that every even number ngreater than 2is the sum of two of them(Goldbach’s Conjecture). This has been directly verified for nup to 10^{18}. We might then hope for a proof along the lines above for Bertrand’s Postulate, where we show that beyond a certain specified integer N, we may introduce a comparison based on what is known about the distribution of the primes to ensure that

there will always be at least one solution in primes p, q to the equation p+q=2n for any even number2n⩾N. This still eludes us, although there are weaker results along these lines- for example, it has been known since 1939 that every sufficiently large odd number is the sum of at most three prime numbers and that every even number is the sum of no more than 300,000 primes. Proof of the full Goldbach Conjecture still seems a long way off.

A simple result that has something of theflavour of the argument type referred to above is that there is a number nless than 4 billion that can be written as the sum of four different cubes in ten distinct ways. It is known that $1729=1^3+12^3=9^3+10^3$ is the smallest number that is the sum of two cubes in two different ways. However, we do not necessarily have to identify the number nin order to know that it must exist. Sometimes it is possible to know for certain that there are solutions to a problem, without actuallyfinding any of those solutions explicitly.

In this case, we begin by noting that if we take four different numbers no more than afixed integer mand form the sum of their cubes, the result is less than 4m3. However, ifm=1000, then an elementary calculation shows that the number of sums of four different cubes is more than 10 times the number $4m^3$, fromwhich it follows that some number $n\leqslant 4m^3=4,000,000,000$ must be the sum of four cubes in at least ten different ways. The details involve calculations

using binomial coefficients(introduced in Chapter 5)and are not especially difficult.

The global picture of prime distribution is summed up by the observation of the leading 19th-century German mathematician and physicist Karl Friedrich Gauss(1777-1855)that p(n), the number of primes up to the number n, is approximately given by n/lognand that the approximation becomes more and more accurate as n increases. For example, if we take nto be one million, the ratio of n/lognsuggests that, up to that stage, about one number in every 12.7 should be prime. Gauss’s observation, which in detail says something more precise, was not proved until 1896.The logarithm function referred to here is the so-called naturallogarithm, which is not based on powers of 10, but rather on powers of a special number e, which is approximately equal to2.718. We shall hear more of this very famous number e in Chapter 6.

The most celebrated undecided question in number theory is the Riemann Hypothesis, which can only be explained in terms of complex numbers, which we have yet to introduce. However, I mention it here as the object of the question can be reformulated using the uniqueness of prime factorization to involve a certain infinite product featuring all the primes. This leads to an interpretation which says that the Hypothesis implies that the overall distribution of the primes is very regular in that, in the long run, primality will apparently occur randomly. Of course, whether or not a

particular number is a prime is not a random event but what is meant is that primality, in the realm of the very large, takes on the mantle of randomness, with no additional pattern or structure to emerge. Many a number theorist has a heartfelt wish to see this 150-year-old conjecture settled in their own lifetime.

Since they represent so natural a sequence, it is almost irresistible to search for patterns among the primes. There are, however, no genuinely useful formulas for prime numbers. That is to say, there is no known rule that allows you to generate all prime numbers or even to calculate a sequence that consists entirely of different primes. There are some neat formulas but they are of little practical worth, some of them even require knowledge of the prime sequence to calculate their value so that they are essentially a cheat. Expressions such as n2+n+41 are known as polynomials, and this one is a particularly rich source of primes. For example, putting n=1, 7, and 20 in turn yields the primes 43, 107, and 461respectively. Indeed, the output of this expression is prime for all values of nfrom n=0ton=39. At the same time, however, it is clear that this polynomial will let us down when we put n=41, as the result will have 41 as a factor, and indeed it fails forn=40as

$$40^2+40+41=40(40+1)+41=40\times 41+41=(40+1)41=41^2.$$

In general, it is quite straightforward to show that no polynomial of this kind can yield a formula for primes, even

if we allow powers higher than 2 to enter the expression. It is possible to devise tests for primality of a number that can be stated in a few words. However, to be of use they would need to be quicker, at least in some cases, than the direct verification procedure described in Chapter 1. A famous result goes by the name of Wilson's Theorem. Its statement involves the use of numbers called factorials, which we will meet again in Chapter 5. The number $n!$, read 'n factorial', is just the product of all numbers up to n. Forexample, $5!=5\times 4\times 3\times 2=120$. Wilson's Theorem is then a very succinct statement: a number p is prime if and only if p is a factor of $1+(p-1)!$.

The proof of this result is not very difficult, and indeed in one direction it is nearly obvious: if p were composite, so that $p=ab$ say, then since both a and b are less than p, they each occur as factors of $(p-1)!$ and so p is a divisor of this factorial as well. It follows that when we divide $1+(p-1)!$ by p, wewillobtaina remainder of 1. (The case where $a=b$ requires a little more thought.) This is very reminiscent of Euclid's proof for the infinity of primes. It follows that if p is a factor of $1+(p-1)!$ then p must be prime. The converse is a little harder to prove: if p is prime then p is a factor of $1+(p-1)!$. This, however, is the surprising direction of the theorem, although the reader can easily verify particular cases: for example, the prime 5 is indeed a factor of

$$1+4!=1+24=25.$$

What Wilson’s Theorem does is convert the problem of determining whether or not p is prime from a series of division problems(checking division by all primes up to $\sqrt{p}$)into a single division problem. However, the subject of the division, $1+(p-1)!$, is huge even for quite small values of p. Despite being a concise statement, Wilson’s Theorem is of no real use in identifying particular prime numbers. For example, to check that 13 is prime by Wilson would require us to verify that 13 is a factor of $1+12!=479,001,601$. (But by applying the divisibility test for 13in Chapter 1, the reader can check that Wilson was right!)Compare this to the labour involved in simply checking that 13 is divisible by neither 2 nor 3. Although Wilson’s Theorem is not useful in prime verification, it has more than ornamental value and can be used to demonstrate other theoretical results.

As afinal observation, we can exploit factorials, which by design have many factors, in order to prove that no arithmeticsequence of numbers, that is to say one of the form $a, a+b, a+2b, a+3b, \cdots$ can consist only of primes as it is possible to show that the gap between successive primes can be arbitrarily large while the common difference between consecutive members of the previous sequence isfixed at b. To see this, consider the sequence of nconsecutive integers:

$$(n+1)!+2,\ (n+1)!+3,\ (n+1)!+4,\ \cdots,\ (n+1)!+n+1.$$

Each of these numbers is composite, as thefirst is divisible by 2(as each of the terms has 2 as a factor), the second is divisible by 3, the next by 4, and so on up until

thefinal one in the list, which has n+1 as a factor. We therefore have, for any given n, a sequence of n consecutive numbers, none of which are prime.

Instead of focusing on numbers with the fewest possible factors(the primes), we shall in the next chapter turn to numbers with many factors, although we shall discover that here too there are surprising links to some very special prime numbers.

Chapter 3 Perfect and not so perfect numbers

Perfection in a number

It is often easy tofind peculiar properties of small numbers that characterize them- for instance,3 is the only number that is the sum of all the previous numbers, while 2 is the only even prime(making it the oddest prime of all). The number 6 has a truly unique property in that it is both the sum and product of all of its smallerfactors: 6=1+2+3=1× 2× 3.

The Pythagoreans called a number like 6 perfect, meaning that the number is the sum of its proper factors, as we shall call them,which are the divisors strictly smaller than the number itself. This kind of perfection is indeed very rare. Thefirstfive perfect numbers are 6,28,496,8128, and 33,550,336. A lot is known about the even perfect numbers but, to this day, no one has been able to answer the basic question of the Ancients as to whether there are infinitely many of these special numbers. What is more,no one has found an odd one, nor proved that there are none.Any odd perfect number must be extremely large and there is a long list of special properties that such a number must possess in consequence of its odd perfection. However, all these restrictions have not as yet legislated such a number out of existence- conceivably, these special properties serve to

direct our search for the elusivefirst odd perfect number, which may yet be awaiting discovery.

The even perfects were known to Euclid to have a tight connection with a very special sequence of primes, known to us as theMersenne primes named after Marin Mersenne(1588-1648), a 17th-century French monk.

A Mersenne number mis one of the form 2p−1, where p is itself a prime. If you take, by way of example, thefirst four primes, 2, 3, 5, and 7, thefirst four Mersenne numbers are seen to be: 3, 7, 31, and127, which the reader can quickly verify as prime. If p were not prime, suppose p=ab say, then $m=2^p-1$ is certainly not prime either, as it can be verified that in these circumstances the number mhas 2^a-1 as a factor. However, if p is prime then the corresponding Mersenne number is often a prime, or so it seems.

And Euclid explained, back in 300 BC, that once you have a prime Mersenne number then there is a perfect number that goes with it, that number being $P=2^{p-1}(2^p-1)$. The reader can soon verify that thefirst four Mersenne primes do indeed give thefirst four perfect numbers listed above: for example, using the third prime 5 as our seed we get the perfect number $P=2^4(2^5-1)=16\times 31=496$, the third perfect number in the previous list.(The factors of P are the powers of 2 up to 2^{p-1}, together with the same list of numbers multiplied by the prime 2^p-1. It is now an exercise in summing what are known

as geometric series(explained in Chapter 5)to check that the proper factors of P do indeed sum to P.)

What is more, in the 18th century the great Swiss mathematician Leonhard Euler(1707- 83)(pronounced ‘Oiler’)proved the reverse implication in that every even perfect number is of this type. In this way, Euclid and Euler together established a one-to-one match between the Mersenne primes and the even perfect numbers. However, the next natural question is, are all the Mersenne numbers prime?Sadly not, and failure is close at hand as thefifth Mersenne number equals $2^{11}-1=2,047=23\times 89$. Indeed, we do not even know if the sequence of Mersenne primes runs out or not- perhaps after a certain point all the Mersenne numbers will turn out to be composite.

The Mersenne numbers are natural prime candidates all the same,as it can be shown that any proper divisor, if one exists, of a Mersenne number mhas the very special form 2kp+1.For example, when p=11, by dent of this result, we need only check for division by primes of the form 22k+1. The two prime factors,23 and 89, correspond to the values k=1andk=4respectively.This fact about divisors of Mersenne numbers also provides a bonus in that it affords us a second way of seeing that there must be infinitely many primes, for it shows that the smallest prime divisor of 2^p-1 exceeds p,andsop cannot be the largest prime.Since this applies to every prime p, we conclude that there is no largest prime and the prime sequence runs on forever.

Since we have no way of producing primes at will, there is, at any one time, a largest known prime and nowadays the champion is always a Mersenne prime, thanks to the international GIMPS venture(Great Internet Mersenne Prime Search). This is a collaborative project of volunteers, which began in 1996. The project uses thousands of personal computers working in parallel, which test Mersenne numbers for primality using a specially devised cocktail of tailor-made algorithms. The current world champion, announced in August 2008, is $2^p-1$where p=43,112,609, although a new Mersenne prime was found in April 2009 with p=42,643,801. These numbers have about 13million digits and would take thousands of pages to write down in ordinary base ten notation.

Less than perfect numbers

Traditional number lore often focused on individual numbers thought to have special, if not magical, properties such as those that are perfect. However, a number pair with a similar trait is220and284, thefirst amicable pair, meaning that the proper factors of each sums to the other- a kind of perfection extended to a couple. The reknowned amateur French mathematician Pierre de Fermat(1601- 65)found other amicable pairs, such as 17,296and 18,416, while Euler discovered dozens more. Surprisingly, they both missed the small pair of 1184 and 1210, found by 16-year-old NicolòPaganini in 1866. We can of course try to go beyond pairs and look for perfect triples, quadruples, and so on. Longer cycles are rare but do crop up.

We can begin with any number, find the sum of its proper divisors, and repeat the process, forming what is known as the number' Saliquot sequence. The result is often a little disappointing in that typically we get a chain that heads to 1 quite rapidly, at which point the process stalls. For example, even beginning with a promising-looking number such as 12, the chain is short:

$$12 \to (1+2+3+4+6) = 16 \to (1+2+4+8) = 15 \to (1+3+5) = 9$$

$$9 \to (1+3) = 4 \to (1+2) = 3 \to 1.$$

The trouble is, once you hit a prime, you arefinished. The perfect numbers are of course exceptions, each giving us a little loop, while an amicable pair leads to a two-cycle: 220→284→220→・・・. Numbers that lead to cycles longer than two are called sociable. They were not studied at all until the20th century as no one had ever found any. Even today, no number that leads to a three-cycle has been discovered, although there are now 120 known cycles of length four. Thefirst examples were found by P. Poulet in 1918. Thefirst is afive-cycle:

$$12,496 \to 14,288 \to 15,472 \to 14,536 \to 14,264 \to 12,496.$$

Poulet' s second example is quite stunning, and to this day no other cycle has been found that comes close to matching it: starting with14,316 we obtain a cycle of length

28. All other known cycles have length less than 10. To the present day, there are no theorems on amicable and sociable numbers as beautiful as those of Euclid and Euler on perfect numbers. However, modern computing power has led to something of an experimental renaissance in this kind of topic and there is more that can be said.

We can divide all numbers into three types, deficient, perfect,and abundant according to whether the sum of their proper divisors is less than, is equal to, or exceeds the number itself. For example, as we have already seen,12 is an abundant number, as are 18 and 24as the respective sums of their proper divisors are 21 and 36.

A naive search for abundance among the integers might lead you to guess that the abundant numbers are simply the multiples of 6.Certainly, any number greater than 6 of the form 6n is abundant,as the factors of 6nmust include 1,2,3 together with n,2n,and3n, which sum to more than the original number 6n.This observation can be extended, however, to show that abundance is not just about sixes as we can argue the same way for any perfect number k.Thefactorsofnk will include 1 together with all the factors of the perfect number k, each multiplied by nso that the sum of all the proper factors of nk will add up to at least 1+nk,and therefore any multiple of a perfect number will be abundant.Forexample,28isperfectandhence2× 28=56,3× 28=84 etc.are all abundant.

And so we see that multiples of perfect numbers and indeed, by the same token, multiples of abundant numbers are themselves abundant. Having made this discovery, you still might guess that all abundant numbers are simply multiples of perfect numbers.However, you don’t have to look too much further tofind thefirst exception to this conjecture, for 70 is abundant but none of its factors are perfect. Indeed,70 is thefirst so-called weird number,but not exactly for this reason(the source of this label is explained below).

Despite these discoveries, you might still think it likely that, just as there seem to be no odd perfect numbers, there are no odd abundant numbers either. In other words, our modified conjecture might be that all odd numbers are deficient. Calculation of the aliquot sums of thefirst few hundred odd numbers would seem to confirm this theory, but the claim is eventually debunked upon testing 945, which has 975 as the sum of its proper divisors. Now the floodgates open as any multiple of an abundant number is abundant, and in particular the odd multiples of 945 immediately supply us with infinitely many more odd abundant numbers.

If we act a little more shrewdly, however, we can discover this counter-example more quickly than if we unthinkingly test one odd number after another. For a number to have a large aliquot sum, it needs lots of factors and large factors at that, which themselves come from being paired with small factors. We can therefore build numbers with large aliquot sums by multiplying small primes together.

If we are focusing on odd numbers only, we should look at those that are products of thefirst few odd primes, which are 3, 5, 7, etc. This rule of thumb would soon lead you to test $3^3 \times 5 \times 7=945$ and thereby discover the abundance property among the odd numbers also.

It is not that unusual tofind that the smallest example of a number with certain properties turns out to be rather large. This is especially true if the specified properties implicitly build a certain factor structure into the required numbers. The smallest example can then turn out to be gigantic, although not necessarily hard to find if we exploit the given properties in our quest for the solution. An example of a number riddle of this kind is tofind the smallest number that isfive times a cube and three times afifth power. The answer is

$$7,119,140,125=5\times 11253=3\times 75^5.$$

The reason why the smallest solution is in the billions, however, is not hard to see. Any solution nhas to have the form 3r5smfor some positive powers r and s and where the remaining prime factors are collected together into a single integer mthat is not divisible by 3 or 5. If wefirst focus on the possible values of r, we observe that since n is 5 times acube, the exponent r must be a multiple of 3, and since nis 3 times a 5th power, the number r-1has to be a multiple of 5. The smallest r that satisfies both these conditions simultaneously is r=6. In the same way, the exponent s has to

be a multiple of 5, while s−1 has to be a mutiple of 3 and the least s thatfits the bill is s=10. Tomakenas small as possible, we take m=1andson=$3^6 \times 5^{10} = 3(3 \times 5^2)^5 = 3 \times 75^5$, sothatn s indeed 3 times a 5th power and at the same time n=$5(3^2 \times 5^3)^3 = 5 \times 1125^3$, andsonis also 5 times a cube.

An even more extreme example is the celebrated Cattle Problem, attributed to Archimedes(287- 212 BC), the greatest mathematician of antiquity. It was not solved until the 19th century. The smallest herd of cattle that satisfies all the imposed constraints in the original 44−line poem is represented by a number with over 200,000 digits!

A warning to be gleaned from all this is that numbers do not display their full variety until we move into the realms of the very large. For that reason, the mere fact that there are no odd perfect numbers with fewer than 300 digits does not in itself give grounds for saying that they ‘probably’ do not exist. All the same, it is the case that some leading experts in thefield would be astonished if one ever turned up.

Returning once again to the general behaviour of aliquot sequences, there are still simple questions that may be put that no one can answer. What possibilities are open to aliquot sequences?If the sequence hits a prime, it will immediately terminate thereafter at 1, and cannot do this in any other way. If this does not happen, the sequence could be cyclic and so represent a sociable number. There is, however,

another related possibility that is revealed by calculating the aliquot sequence of 95:

$$95=5\times 19\rightarrow (1+5+19)=25=5\times 5\rightarrow (1+5)=6\rightarrow 6\rightarrow 6\rightarrow \cdots.$$

What has happened here is that although 95 is not itself a sociable number, its aliquot sequence eventually hits a sociable number(or more precisely in this case, the perfect number 6)and then goes into a cycle.

There is conceivably one possibility remaining, that being that the aliquot sequence of a number never hits a prime nor a sociable number, in which case the sequence must be an unending series of different numbers, none of which are either prime or sociable. Is this possible?Surprisingly, no one knows. What is more surprising is that there are small numbers whose aliquot sequence remain unknown(and thereby remain candidates for having such an infinite aliquot sequence). Thefirst of these mysterious numbers is276, whose sequence begins:

$$276\rightarrow 396\rightarrow 696\rightarrow 1104\rightarrow 1872\rightarrow 3770\rightarrow 3790\rightarrow$$
$$\rightarrow 3050\rightarrow 2716\rightarrow 2772\rightarrow \cdots$$

but no one knows exactly where it ends up.

It might well be that the reader would like to explore a little on their own, in which case I should let you in on the secret of how to calculate the so-called aliquot function a(n)fromtheprime factorization of n- take the product of all

terms$(p^{k+1}-1)/(p-1)$, where p^k is the highest prime power of the prime p that divides n, and then subtract n itself. For example, $276=2^2\times 3\times 23$ and so as indicated by the second term in the aliquot sequence for 276listed above.

$$a(276) = \frac{2^3-1}{2-1} \times \frac{3^2-1}{3-1} \times \frac{23^2-1}{23-1} - 276 =$$
$$7 \times 4 \times 24 - 276 = 672 - 276 = 396$$

There is no end to the types of numbers that we can introduce by giving a name to the numbers nthat bear a certain relationship to the aliquot function. As we have already mentioned, nis perfect if a(n)=nand abundant if a(n)> n. Asemiperfect number n is one that is the sum of some of its proper divisors(those less than n), so it follows from the definition that all semiperfect numbers are either perfect or abundant. For example, 18 is semiperfect as18=3+6+9. A number is called weird if it is abundant but not semiperfect, and the smallest weird number is 70.

One can take the view that the topic is becoming too miscellaneous in nature- bestowing names on rather arbitrarily defined classes of numbers does not of its own accord make them interesting. We should know when to stop. That said, it is worth appreciating that the underlying strategies used to tackle these new questions are yet reminiscent of what Euclid and Euler showed us in relation to perfect numbers. You will recall that what Euclid proved was

that if a Mersenne number was prime then another number was even and perfect. Euler then proved conversely that all even perfect numbers arise from this approach. In the 9th century, the Persian mathematician Thabit ibn Qurra introduced for any number na triple of numbers which, if all prime, allowed the construction of an amicable pair. Thabit' s construction was generalized further by Euler in the 18th century, but even this enhanced formulation only seems to yield a few amicable pairs and there are many amicable pairs that do not arise from this construction.(There are now nearly 12 million known pairs of amicable numbers.)In modern times, a similar approach by Kravitz gives a construction of weird numbers from certain numbers should they happen to be prime,and this formula has successfully found a very large weird number with more thanfifty digits.

These last two chapters have served to familiarize the reader with factors and factorization of the natural numbers, or positiveintegers are they are also known, illustrated through a variety of examples. This will stand you in good stead for the upcoming chapter, in which you will learn how those ideas are applied to contemporary cryptography, the science of secrets.

Chapter 4 Cryptography:the secret life of primes

The reader will now appreciate that the collection of counting numbers has, from the earliest times, been recognized as the repository of riddles and secrets, many of which have never been revealed to this day. For many of us, this is enough to justify the continued serious study of numbers but others may take a different attitude. Intriguing and difficult as these conundrums may be, it might be imagined that they have little bearing on the rest of human wisdom. But that would be a mistake.

Over the last few decades it has emerged that ordinary secrets, of thekindsweallindulgeinfromtimetotime,canbecodedas secrets about numbers. This has now all been put into practice and our most precious secrets, whether they be commercial or military, personal orfinancial, political or downright scandalous,can all be protected on the Internet by masking them using secrets about ordinary counting numbers.

Secrets turned into numbers

How is all this possible?Any information, whether it be a poem or a bank statement, a blueprint for a weapon or a computer program, can be described in words. We may, however, need to augment the alphabet that is used to make up our words beyond the ordinary letters of the alphabet. We may include number symbols, punctuation symbols including special

symbols for space between ordinary words, but it is nonetheless the case that all the information we wish to transfer, including instructions for producing pictures and diagrams, can be expressed using words from an alphabet of, let us say, no more than one thousand symbols. We can count these symbols and so represent each symbol uniquely as a number. Since numbers are cheap and inexhaustible, it may be convenient to use numbers all with the same number of digits for this purpose(so, for example, every symbol was represented uniquely by its own four-digit PIN). We could string the symbols together as required to give one big long numberthattoldtheentirestory.Wecanevenworkinbinaryif we wish and so devise a way of translating any information into one long string of 0s and 1s. Every message we might ever want to send could then be coded as a binary string and then decoded at the other end by a suitably programmed computer, to be compiled in ordinary language that we can all comprehend. This then is the first realization: in order to send messages between one person and another, it is enough, both in theory and in practice, to be able to send numbers from one person to another.

Turning messages into numbers, however, is not the big idea. To be sure, the exact process by which all the information is digitized may be hidden from the general public, but nonetheless is not the source of protection from eavesdroppers. Indeed, from the point of view of cryptography, we may identify any message, the so-called plaintext, with the number that represents it and thereby

think of that number as the plaintext itself, as it is assumed that anyone has access to the wherewithal that will allow one to be transformed into the other. Secrecy only comes on to the scene when we mask these plaintext numbers with other numbers.

Let me introduce you to thefictitious characters that populate the various scenarious of cryptography, which is the study of ciphers(secret codes). We imagine Alice and Bob, who want to communicate with each other, without being overheard by the eavesdropper, Eve. Instinctively, we might sympathize with Alice and Bob, picturing Eve as up to no good, but of course the reverse may be true, with Eve representing a noble policing authority striving to protect us all from the evil plots of Bob and Alice.

Whatever the moral standing of the participants, there is an age-old approach that Alice and Bob may employ to cut Eve out of the conversation even if Eve intercepts messages that pass between them. They can encrypt the data using a cipher key that is known only to Alice and Bob. What they may arrange to do is to meet in a secure environment where they exchange with one another a secret number(let us say 57)and then return home.When the time comes, Alice will want to send a message to Bob and, just to illustrate the point, suppose that message can be represented by a single digit between 1 and 9. On the big day, Alice wants to send the message ‘8’ to Bob. She takes her message and adds the secret ingredient, that is to say she masks its true value by adding 57 and so

sends the message to Bob, across an insecure channel, of 8+57=65. Bob receives this message and subtracts the secret number to retrieve Alice’s plaintext 65-57=8.The nefarious Eve, however, has a good idea what these two are up to and indeed does manage to intercept the enciphered message,65. But what can she do with it?She may know, as we do, that Alice has sent one of the nine possible messages 1,2,3,・・・,9to Bob and also knows that she has encrypted it by adding a number to the message, which must therefore lie between 65-9=55and65-1=64. However, because she cannot tell which of these nine masking numbers has been used(she isn’t in on the secret), she is none the wiser as to the actual plaintext message that Alice sent to Bob, which is still just as likely to be any one of the nine possibilities. All she knows is that Alice has sent a message to Bob but has no idea what it is.

It might seem that Alice and Bob are now impervious to the malice of Eve and can communicate with impunity using the magic number 57 to disguise all they have to say. That, however,isn’t quite the case. They would be well advised to change that number, indeed they are better off using a new secret number every time because if they don’t, the system will begin to leak information to Eve. For example, say in a future week Alice wants to send to Bob the same message number 8. Everything would run as before and once again Eve would intercept the mysterious number 65 from the airwaves, but this time it would tell her something. Eve would know that, whatever this message is, it is the same message that

Alice sent to Bob in thefirst week- this is just the sort of thing Alice and Bob would not want Eve to know.

This, however, looks to be no big problem for Alice and Bob.When theyfirst meet up to ‘exchange keys’, instead of agreeing on just one secret number, Alice could provide Bob with a long ordered list of thousands of secret numbers, to be used one after another, thus avoiding the possibility of meaningful coincidences in their publicly available communications.

And this is indeed what is done in practice. This kind of cipher system is known in the trade as a one-time pad.Thesenderand receiver mask their plaintext with a single-use number from the ‘pad’. That leaf of the pad is then discarded by both the sender and receiver after the message has been sent and deciphered. The one-time pad represents a completely secure system in that the insecure message that travels in the public domain contains no information about the content of the plaintext. To decipher it, the interceptor needs to get hold of that pad in order to obtain the encryption- decryption key.

Keysandkeyexchange

It would seem then that the problem of secure communication is completely solved by the one-time pad and, in a way, that is true.The difficulty with ciphers like the one-time pad, however, is that they require the participants to exchange a key in order to use them. In practice, this

takes a lot of effort. For high-level communications, such as those between the White House and the Kremlin, money is no object and the necessary exchanges are carried out under conditions of maximum security. In the everyday world on the other hand, all sorts of people and institutions need to communicate with one another in a confidential fashion. The participants cannot afford the time and energy required to secure key exchanges and, even if this were arranged by a trusted third party, it can be an expensive business.

The common drawback of all ciphers that had been used for thousands of years up until the 1970s was that they were allsymmetric ciphers, meaning that the encryption and decryption keys were essentially the same. Whether it was the simple alphabet-shift cipher of Julius Caesar, or the complex Enigma Cipher of the Second World War, they all suffered from the common weakness that once an adversary learned how you were encoding your messages, they could decode them just as well as you. In order to make use of a symmetric cipher, the communicating partners needed to exchange the cipher key in a secure way.

It seemed to have been tacitly assumed that this was an unavoidable principle of secret codes- for a cipher to be used the partners needed, somehow or other, to exchange the key to the cipher and to keep it secret from the enemy. Indeed, this might be regarded as mathematical common sense.

This is the kind of assumption that makes a mathematician suspicious. We are dealing with what is essentially a mathematical situation, so one would expect such a ‘principle’ to be well founded and represented by some form of mathematical theorem. Yet there was no such theorem, and the reason that there was no such theorem was that the principle simply is not valid, as the following thought experiment reveals.

Transmission of a secure message from Alice to Bob does not in itself necessitate the exchange of the key to a cipher, for they can proceed as follows. Alice writes her plaintext message for Bob, and places it in a box that she secures with her own padlock. Only Alice has the key to this lock. She then posts the box to Bob, who of course cannot open it. Bob, however, then adds a second padlock to the box, for which he alone possesses the key. The box is then returned to Alice, who then removes her own lock, and sends the box for a second time to Bob. This time, Bob may unlock the box and read Alice’ s message, secure in the knowledge that the meddling Eve could not have peeked at the contents during the delivery process. In this way, a secret message may be securely sent on an insecure channel without Alice and Bob ever exchanging keys. This imaginary scenario shows that there is no law that says that a key must change hands in the exchange of secure messages. In a real system, Alice and Bob’ s ‘locks’ might be their own coding of the message rather than a physical device separating the would-be eavesdropper from the plaintext. Alice and Bob may then use

this initial exchange to set up an ordinary symmetric cipher that would be used to mask all their future communication.

Indeed, this is the way a secure communication channel is often established in the real world. Replacing physical locking devices by personal codes is not, however, so easy to do. Unlike the locks, the encodings of Alice and Bob may interfere with one another, making the unscrambling(that is, the unlocking)that is carried outfirst by Alice and then by Bob unworkable. However, that this method can be effective wasfirst publicly demonstrated by Whitfield Diffie and Martin Hellman in 1976.

A second related approach is the idea of asymmetric or public key cryptography in which everyone publishes their own public key that is then used to encipher messages meant for that person. However, each person also holds a private key, without which the messages enciphered using their own unique public key cannot be read. In terms of the padlock metaphor, Alice provides Bob with a box in which to place his plaintext message together with an open padlock(her public key)to which she alone holds the key(her private key).

A workable public key system might seem too much to ask for as the twin requirements of security and ease of use seem to conflict. Fast, safe encryption is, however, available to the general public on the Internet, even if they barely realize that it is there, safeguarding their interests. And it is all down to numbers, and prime numbers at that.

How secret primes protect our secrets

Remember that every plaintext message is regarded just as a single number, so it is natural to try to mask this number using other numbers. The most common way to do this is through employing the so-called RSA enciphering process, published in 1978 by its founders, Ron Rivest, Adi Shamir, and Leonard Adleman. In RSA, each person's private key consists of three numbers, p, q, andd, where p and q are(very large)prime numbers and the third ingredient d is Alice's secret deciphering number, the role of which will be explained in due course. Alice provides the public with n=pq, the product of her two secret primes, and an enciphering number e(which is an ordinary whole number, in no way related to the special constant called e mentioned in Chapter 2).

A simple example for the purposes of illustration would be for Alice to have the primes p=5andq=13sothatn=5× 13=65. If Alice sets her enciphering number to be e=11, then her public key would be(n, e)=(65, 11). To encrypt a message m, Bob only needs nand e. However, to decipher the encrypted message E(m)that Bob transmits to Alice requires the deciphering number d, which in this case turns out to be d=35, as we shall show a little further on. The mathematics that allows d to be calculated requires that the primes p and q are known. In this toy example, given that n=65, anyone would soon discover that p=5and q=13. However, if the primes p and q are extremely large(typically they are hundereds of digits in length), this task becomes a practical impossibility for almost any

computer system, at least in a reasonably short time, such as two or three weeks. In summary, the RSA system of enciphering is based on the empirical fact that it is prohibitively difficult tofind the prime factors of a very, very large number n. The clever part, which we shall explain in the remainder of the chapter, lies in devising a way that the message number mcan be enciphered just using the publicly known numbers nand e but, in practice, deciphering requires possession of the prime factors of n.

Here is how it works. What Bob sends through the ether is not m itself but the remainder when m^e is divided by n. Alice can then recover mby taking this remainder r and similarly calculating the remainder when r^d is divided by n. The underlying mathematics ensures that the outcome for Alice is the original message m, which can then be decoded into ordinary plaintext by Alice's computer system. This is, of course, happening seamlessly behind the scenes for any real-life Alice and Bob.

It would seem that the only thing that Eve lacks that really matters is this deciphering number d. If Eve knew that, she could decipher the message just as well as Alice. It turns out that d is a solution of a certain equation. Solving this equation is computationally quite easy and relies on the Euclidean Algorithm, published in the Books of Euclid in 300 BC. That is not the difficulty. The trouble is that it is not possible tofind out exactly what equation to

solve unless you know at least one of the primes p and q, and that is the obstacle that stops Eve in her tracks.

We can explain more about how the numbers involved in all this work into the system. First, there is apparently quite a problem with Bob’ s initial task. The number mis big, the number nis monstrous(of the order of 200 digits)and even if e is not that large, the number me is going to be extremely large as well. After calculating it, we have to divide me by the number nto get the remainder r, which represents the enciphered text. It might seem that the calculations are too unwieldy to be practical. We should be aware that even though modern computers are extremely powerful, they yet have their limitations. When calculations involve very high powers, they can exceed the capacity of any computer system. We certainly cannot assume that any practical calculation that we set for a computer can be done in a short period of time.

The saving grace for Bob is that it is possible tofind the required remainder r, without doing the long division at all. Indeed, the remainders just depend on remainders, and here is an example to illustrate the point. What are thefinal two digits of 7^{39}?(Thatisto say, what is the remainder when this number is divided by 100?)In order to answer this question, we might begin by calculating thefirst few powers of 7: $7^1=7$, $7^2=49$, $7^3=343$, $7^4=2,401$, $7^5=16,807$, • • •. It will soon become clear, however, that the sheer size of these numbers is going to become unmanageable well before we get anywhere near 739. On the other hand, as we calculate one power after

another, we see a pattern emerging. The key observation is that, as we calculate succeeding powers, thefinal two digits of the answer depend only on thefinal two digits of the preceding number, as when we carry out the multiplication, digits in the hundreds column and beyond have no effect on what end up in the units and tens columns.

What is more, since 74 has 01 as thefinal digit pair, the next four powers will end in 07, 49, 43, and then 01 again. Hence, as we compute succeeding powers, the pattern of the last two digits will simply repeat this cycle of length four, over and over again. To return to the question in hand, since $39=4\times 9+3$, we will pass through this four-cycle nine times and then take three more steps in calculating thefinal two digits of 7^{39}, which must therefore be 43.

And this works quite generally. In order tofind the remainder when some power a^b is divided by nsay, we need only take the remainder r when a is divided by nand keep track of the remainders as we take successive powers of r.Whenweworkwith the remainder r, which will be a number in the range from 0 to n-1, mathematicians say that we are working modulo n,discarding any higher multiples of nthatmayarise,astheyleavea remainder of 0 when divided by n, and so cannot contribute to the value of thefinal remainder r.

You might still suspect that I have rigged the evidence by choosing an example where a very small power left a

remainder of 1- in this instance 74 was 1 more than a multiple of n=100. This, however, is only partly true. It turns out that, if we take any two numbers, a and n, whose highest common factor is 1(we say such numbers are mutually coprime), then there is always a power t such that a^tequals 1 modulo n, that is to say leaves a remainder of 1 when divided by n. From this point, the remainders of successive powers follow a cycle of length t. Itcan, however, behardtopredictwhat is the value of t, but it is known that t must always equal or be a factor of a number traditionally written asϕ(n), the value of the Euler phi function.

And so what isϕ(n)?It is defined as the count of the numbers up to nthat are coprime with n. For example, if n=15, then the set of numbers in question is{1, 2, 4, 7, 8, 11, 13, 14}, which lists all the numbers up to 15 that have no factor in common with 15(except the inevitable factor of 1). Since this set has 8 members, we see thatϕ(15)=8. Fortunately there is a slicker way offinding?(n)that does not entail explicitly listing all the integers coprime with n and then counting them up. As with most functions of this nature, the value can be expressed in terms of the prime factorization of n. Indeed, we only need know the prime factors of n, forϕ(n)canbe found by taking nand multiplying it by each of the fractions($\frac{1}{p}$)asp ranges over all the prime divisors of n. Theprime factors of 15, for instance, are 3 and 5, so the answer in this case is which is the same as the result that we obtained directly from the definition. Using this method, you might like to check yourself thatϕ(100)=40, and so, for instance, it then

follows that 7^{40}equals 1 modulo 100. However, as we have already seen, the least power of 7 that yields a remainder of 1 is not 40 but its divisor 4.

$$\phi(15) = 15 \times (1 - \frac{1}{3})(1 - \frac{1}{5}) = 15 \times \frac{2}{3} \times \frac{4}{5} = 8,$$

All this serves to give an indication that the number sent by Bob to Alice, m^e modulo n, can indeed be calculated without too much effort on behalf of Bob’ s computer. All the same, the numbers involved are in practice mighty big, so more explanation is needed to show that they can be handled. The large powers involved in computing me can be dealt with in stages by a process known asfast exponentiation. Without going into detail, the method involves successive squaring and multiplying of powers to arrive at m^e modulo n with the binary form of e guiding the algorithm through to quicklyfind the required remainder in relatively few steps.

Euclid shows Alice how tofind her deciphering number

The deciphering number is the receiver’ s magic wand that allows the message retrieval. This number d is chosen so that the product de leaves a remainder of 1 when divided byϕ(n). Sinceϕ(n)=pq is a product of two distinct primes, the value of It turns out that there is always just one value for d in the range up toϕ(n)tha tha sthe required property.

Alice's computer can find d using an algebraic tool that is over 2,300 years old, the Euclidean Algorithm, which will be explained in a moment. Eve's computer could of course do the same thing if it just knew which equation to solve. However, since p and q are private to Alice, so is (p-1)(q-1) and Eve does not know where to begin.

The existence of d is only guaranteed if a certain mild restriction is placed on the (publicly declared) enciphering number e. Alice must ensure that e has no common prime factor with $\phi(n)$. This is quite easily done as Alice can test $\phi(n)$ for division by particular primes and so ensure that e meets these requirements without compromising the identity of p and of q. Indeed, the value of e often used in practice is the fourth so-called Fermat prime $e=65537=2^{16}+1$; this value, $2^{24}+1$ has a particularly rare property, that being that it is possible to construct a regular polygon with e sides using a straightedge and compass. Its utility in cryptography, however, is due to it being a fairly large prime that exceeds a power of 2 by exactly 1, which lends well to the fast exponentiation process mentioned earlier.

Returning to the Euclidean Algorithm, this begins from the observation that it is possible to find the highest common factor (hcf) of two numbers $a>b$ by successive subtraction. (The hcf is also known as the gcd- greatest common divisor.) We just note that $r=a-b$ has the property that any common factor of any two of the three numbers a, b, and r will also be a factor of the third. For example, if c is a common factor of a and

b, sothata=ca_1 and b=cb_1 say, we see that $r=a-b=ca_1-cb_1=c(a_1-b_1)$, giving us a factorization of r involving the divisor c. In particular, the hcf of a and b is the same as the hcf of b and r. Since both these numbers are less than a, we now have the same problem but applied to a smaller number pair. Repetition of this idea then will eventually lead to a pair where the hcf is obvious. (Indeed, the two numbers in hand will eventually be the same, for if not we could proceed one more step; their common value is then the number we seek.)

Forexample, tofindthehcfofa=558 and b=396, thefirst subtraction would give us r=558−396=162, so our new pair would be 396 and 162. Since 396−162=234, our third pair becomes 234 and 162, and as we continue the full list of number pairs is:

$$(558, 396) \to (396, 162) \to (234, 162) \to (162, 72) \to (90, 72) \to$$
$$\to (72, 18) \to (54, 18) \to (36, 18) \to (18, 18)$$

and so the hcf of 558 and 396 is 18.

It is possible to write down the hcf of a number pair from the prime factorizations of the numbers in question. In this example, $558=2\times 3^2\times 31$, while $396=2^2\times 3^2\times 11$; taking the common power of each prime entering into the factorizations, we obtain the hcf as $2\times 3^2=18$. Nevertheless, for larger numbers it takes much less work to use Euclid' s Algorithm as it is generally easier to perform subtractions than tofind prime factorizations.

Another bonus of the Euclidean Algorithm is that it is always possible to work it backwards and in so doing express the hcf in terms of the original two numbers. To see this in action in the previous example, it is best to compress the calculation when the same number appears several times over in the course of the subtractions, representing this as a single equation as follows:

$558=396+162$

$396=2\times 162+72$

$162=2\times 72+18$

$72=4\times 18.$

Beginning with the second to last line, we now use each little equation to eliminate the intermediate remainders, one at a time. In this example, by usingfirst the penultimate equation, and then the one above that we obtain:

$$18=162-2\times 72=162-2\times (396-2\times 162)=5\times 162-2\times 396$$

andfinally using thefirst equation we can eliminate thefirst intermediate remainder of 162:

$$=5\times (558-396)-2\times 396=5\times 558-7\times 396=18.$$

That we can perform this reverse procedure is important for both practical and theoretical reasons. In particular, tofind Alice’s deciphering number d, wewantd to satisfy the condition that de leaves a remainder of 1 when divided by ϕ

(n).(For brevity, we shall denote ϕ(n)by the single symbol k.)We can now see the reason why we insist on e and kbeing a coprime pair, as if their highest common factor is 1, when we act the Euclidean Algorithm on the pair e and k, thefinal remainder that appears is, of course,1. By reversing the algorithm, we will eventually express 1 as a combination of e and k;in particular, we willfind integers c and d such that ck+de=1, or in other words de=1-ck,sothatde will leave a remainder of 1 when divided by k.

This relatively simple process will yield Alice's deciphering number d:the initial value of d obtained from the equation may not lie in the range from 1 to k but if not, by adding or subtracting a suitable multiple of k, we will eventuallyfind the unique number d in that range that has the magic property that de leaves a remainder of 1 when divided by k.(The uniqueness of d is easily proved, but we won't digress into further explanation here.)That is how the deciperhing number d is calculated as we can show by returning to the example given earlier where p=5,q=13,sothat n=pq=5× 13=65. We have ϕ(n)=(p-1)(q-1)=4× 12=48.Alice sets e=11, and since 11 and 48 are coprime, this is within the rules of the game. The Euclidean Algorithm applied to ϕ(n)=k=48ande=11thengives:

48=4× 11+4

11=2× 4+3

$4=1\times 3+1$ confirming that the hcf of k and e is indeed 1. Reversing the algorithm we obtain:

$$1=4-3=4-(11-2\times 4)=3\times 4-11=3(48-4\times 11)-11=3\times 48-13\times 11.$$

This gives an initial value of $d=-13$ as the solution to the requirement that 11d leaves remainder 1 upon division by 48, so in order to get a positive value of d in the required range we add 48 to this number to get $d=48-13=35$.

The reason why d works for Alice is all down to modular arithmetic and the fact that de leaves a remainder of 1 when divided by $k=\phi(n)$. Alice calculates $(m^e)^d=m^{de}$ modulo n. Now de has the form $1+kr$ for some integer r. As explained before, m^k leaves a remainder of 1 when divided by n (this is often known as Euler's Theorem) and so the same is true of $(m^k)^r=m^{kr}$. Hence $m^{1+kr}=m\times m^{kr}$ leaves the remainder m when divided by n. (Detailed verification of this requires a little algebra, but that is what happens.) In this way, Alice retrieves Bob's message, m.

And in passing it is well to point out that the Euclidean Algorithm provides the missing link in our proof of the uniqueness of prime factorization as it allows us to verify the euclidean property that if a prime p is a factor of the product ab, so that ab=pc say, then p is a factor of at least one of a and b. The reason for this is that if p is not a factor of a then, since p is prime, the hcf of a and p is 1. By reversing the Euclidean Algorithm when applied to the pair a and p, we can then find integers r and s say such that

ra+sp=1. This is enough to show that p is then a factor of b for, since ab=pc, we have:

$$b=b\times 1=b(ra+sp)=r(ab)+psb=r(pc)+psb=p(rc+sb).$$

This is the required factorization of b that features the prime p as afactor.

In conclusion, the number theory underlying RSA enciphering makes the system sound, although various protocols that have not been explained here must be respected in order to safeguard the integrity of the system. There are issues of authentification (what if Eve contacts Alice pretending to be Bob?), non-repudiation (what if Bob pretends that it was Eve who sent his message to Alice?), and identity fraud (what if Alice abuses confidential identification sent to her by Bob and tries to impersonate him online?). Moreover, other weaknesses in the system can be exposed when predictable or repeated messages proliferate. However, these difficulties may potentially arise in any public key cryptosystem. They can be overcome and in the main are unrelated to the underlying mathematical techniques that ensure high quality and robust encyryption.

This chapter has demonstrated a major application of prime numbers and the theory of divisibility and remainders. The ancient mathematics of Euclid and the 18th-century contribution of Euler allows this to be explained, not only in broad principle, but infine detail.

The first part of our book closes with Chapter 5 which introduces some special classes of integers associated with the enumeration of some naturally occurring groupings.

Chapter 5 Numbers that count

Numbers that arise of their own accord in counting problems are important and so have been extensively investigated. Here I will describe the binomial coefficients, and the numbers of Catalan, Fibonacci, and Stirling because they enumerate certain natural collections. But wefirst begin with some very fundamental number sequences.

Triangular numbers, arithmetic and geometric progressions

Since they will reappear when we look at binomial coefficients, I will take a moment to revisit the triangular numbers, thenth of which, denoted by tn, isdefinedasthesumofthefirstncounting numbers. Its value, in terms of n, can be found by the following trick. We write tn as the sum just mentioned and then again as the same sum but in the reverse order. Adding the two versions of tn together:

$$t_n=1+2+3+\cdot\cdot\cdot+(n-2)+(n-1)+n$$

$$t_n=n+(n-1)+(n-2)+\cdot\cdot\cdot+3+2+1;$$

the outcome is of course an expression for 2tn. However, the point of doing this is that we have paired 1 with n, 2withn-1, 3 with n-2, etc. The sum of each of these pairs is the same value, n+1, and there are n pairs altogether. In conclusion, we infer that$2t_n=n(n+1)$, or in other words, the

value of the nth triangular number is $\frac{1}{2}n$(n+1). For example, the sum of all the integers from1 up to 1000 is therefore equal to 500× 1001=500,500.

Indeed, this formula allows us tofind the rule for summing the first nterms of any arithmetic series, or progression as they are known, which is one of the form a, a+b, a+2b, a+3b, •••.We first take care of the change of scale. By multiplying through the expression for tn by b we see that b+2b+3b+•••+nb=$\frac{b}{2}n$(n+1).Tofind the formula for the sum of the general arithmetic series, first note that the sum of thefirst n−1 terms of the previous series is obtained by replacing nby n−1 in the previous formula, to give$\frac{b}{2}n$(n−1). The general arithmetic series now comes from adding a to every term and including a as thefirst term as well. This means that we need to add na to the previous sum to give us the general formula for the sum of thefirst n terms of an arithmetic series:

$$a + (a + b) + (a + 2b) + \cdots + (a + (n-1)b) = na + \frac{b}{2}n(n-1).$$

Forexample, bytakinga=1andb=2, we see that the sum of the first nodd numbers is n+n(n−1)=n+n^2−n=n^2, thenth square.

If we replace addition by multiplication as the operation, we move from arithmetic series to geometric series. In an arithmetic series, each pair of successive terms is separated by a common difference, the number b in our notation.

In other words, to move from one term to the next, we simply add b. In a geometric series, we once again begin with some arbitrary number, a asthefirsttermand move from one term to the next by multiplying by afixed number, called the common ratio, denoted by the symbol r. Thatistosay, the typical geometric series has the form a, ar, ar^2, • • • with the nth term being ar^{n-1}. As with arithmetic series, there is a formula for the sum of thefirst nterms of a geometric series:

$$a + ar + ar^2 + \cdots + ar^{n-1} = \frac{a(r^n - 1)}{r - 1}.$$

The quick way of seeing that this formula is right is to take the equivalent form that we obtain when we multiply both sides of this equation by(r−1)and multiply out the brackets. On the left-hand side we obtain:

$$(ar+ar^2+ar^3+\cdots+ar^n)-(a+ar+ar^2+\cdots+ar^{n-1})$$

and the whole expression telescopes, meaning that nearly every term is cancelled by one in the other bracket: the only exceptions are $ar^n-a=a(r^n-1)$, showing that our formula for the sum is correct. For example, putting a=1andr=2 gives us the sum of powers of 2:

$$1+2+4+\cdots+2^{n-1}=2^n-1.$$

This formula is just what you need in order to verify Euclid’ s result from Chapter 3 on how to generate even perfect numbers from Mersenne primes.

Factorials, permutations, and binomial coefficients

As we have seen, the nth triangular number arises from summing all the numbers from 1 up to n together. If we replace addition by multiplication in this idea, we get what are known as thefactorial numbers, which made theirfirst appearance in Chapter 2.

Factorials come up constantly in counting and probability problems such as the chances of being dealt a certain type of hand in a card game like poker. For that reason, they have their own notation: the nth factorial is denoted by $n! = n\times(n-1)\times\cdots\times 2\times 1$. The triangular numbers grow reasonably quickly, at about half the rate of the squares, but the factorials grow much faster and soon pass into the millions and millions: for example $10! = 3,628,800$. The exclamation mark alerts us to this rather alarming rate of growth.

In particular, $n!$ is the number of recognizably different ways that you can arrange n objects, such as nnumbered balls, in a row. This follows as you have n choices for which ball goesfirst, and for each such choice you have n−1 balls remaining for the second place, n−2 for the third, and so on. If we stop after selecting just r balls, and denote the corresponding number by P(n, r), we see that there are $n\times(n-1)\times(n-2)\times\cdots\times(n-r+1)$permutations, as we say, of n balls, taking r at a time. This is also conveniently expressed as the ratio of two factorials: $P(n, r)=\frac{n!}{(n-r)!}$. For example, if you are dealt a poker hand, you pick up 5 cards from a deck of 52 and

the number of ways this can happen is $P(52,5)=52\times 51\times 50\times 49\times 48$. However, a hand in a card game does not depend on the order in which you pick up the cards but only the collection of cards itself. For a poker hand, each set of5 can itself be rearranged in $5!=120$ ways so that the number of genuinely different 5-card hands is $P(52,5)/120=2,598,960$, about two and a half million.

The most special class that emerges in counting problems, orenumerations as they are called, is that of the binomial coefficients, so named as they arise as the multipliers of powers of x when the binomial expression$(1+x)n$ is expanded. The binomial coefficient $C(n,r)$is the number of different ways we may construct a set of size r from one of size n.Forexample, $C(4,2)=6$, as there are six pairs(taken without regard to order within a pair)that can be chosen from a group of four: for example, if we have four children, Alex, Barbara, Caroline, and David, there are six ways that we can selectapairfromthisgroup: AB, AC, AD, BC, BD, andCD. The binomial coefficients can be calculated in two distinct ways. First, we can extend the argument above that we used to calculate $C(52,5)$: in general we see that $C(n,r)=P(n,r)/r!$, which in turns gives us the useful expression:

$$C(n,r)=\frac{n!}{(n-r)!r!}.$$

A notable special case is when we let r=2, which corresponds to the number of pairs that can be selected from a set of nobjects.The answer is $\frac{n!}{(n-2)!2!}$.Now all the factors in the(n−2)! term in the denominator cancel with the corresponding factors in n! and since 2! =2, the expression for C(n, 2)simplifies to $\frac{n(n-1)}{2}$.Inother words, the number of ways of choosing a pair from a collection of nobjects is t_{n-1}, the(n−1)st triangular number. For instance, as we have already seen, C(4, 2)=6, which is indeed the third triangular number.

This factorial-based formula for calculating binomial coefficients does give a nice algebraic hold on the binomial coefficients that allows us to demonstrate their many special properties. However, the evolution of these properties is often more transparent if we focus on a second way to generate these integers, which is by means of the Arithmetic Triangle(see Figure 2), also known as Pascal’s Triangle, in honour of the 17th-century French mathematician and philosopher Blaise Pascal(1623- 62).(The Arithmetic Triangle has been discovered and re-discovered throughout Persia, India, and China over the last 1,000 years: for example, it featured as the front cover of The Precious Mirror by Chu Shih-Chieh in 1303.)

1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

1 7 21 35 35 21 7 1

1 8 28 56 70 56 28 8 1

⋮

2. The Arithmetic Triangle

Each number in the body of the triangle is the sum of the two above it. The triangle, which can be continued indefinitely, gives the full list of binomial coefficients. For example, tofind the number of ways of selectingfive people from a group of seven,proceed as follows. Number the lines of the triangle, beginning with 0 at the top. Similarly number the positions within each row from left to right, again starting with 0. Go down to the line numbered 7, and then go to the number on that line numbered 5(remembering to start your count from 0):we see the answer is 21.You will note the symmetry of each row:for example,21 is also the number of ways of choosing two people from a group of seven.This is explained by observing that when we choose thefive from seven, we are simultaneously choosing two from seven as well- the two being the pair left behind. This symmetry argument of course applies to every row. This is also manifested in the formula on page 55, for it returns the

same expression if we replace r by n−r, asthetermsr and n−r that we see in the denominator simply swap positions.

The reason that the pattern gives the right answers is not hard to see. Each row builds from the one above it. We can see easily that thefirst three rows are correct: for example, the 2 in the centre of the third row tells us that there are two ways of choosing a single person from a pair. The 1 that sits on top is saying that there is one way to choose a set of size zero from the empty set. In fact, there is one way of choosing a set of size zero from any set, which is why every row begins with 1. Let us focus on the example just given– there are 21=15+6 ways of selectingfive from a group of seven people. The 21 quintets naturally split into two types. First, there are 15 ways to form a group of four from thefirst six people, to which we may add the seventh person to form our fivesome. If we don't include the seventh person, however, then we have to build a set offive from thefirst six, and there are six ways of doing this. This illustrates how one row leads to the next: each entry is the sum of the two above it, and this pattern propagates throughout the triangle. In symbols this rule takes the form:

$$C(n,r)=C(n-1,r)+C(n-1,r-1).$$

The triangle is rich in patterns. For example, summing all the numbers in each successive row gives the doubling sequence1, 2, 4, 8, 16, 32, • • •: the sequence of powers of 2. In summing the row that begins 1, 8, 28, 56, • • • for instance, we are summing the number of ways of choosing a set of size

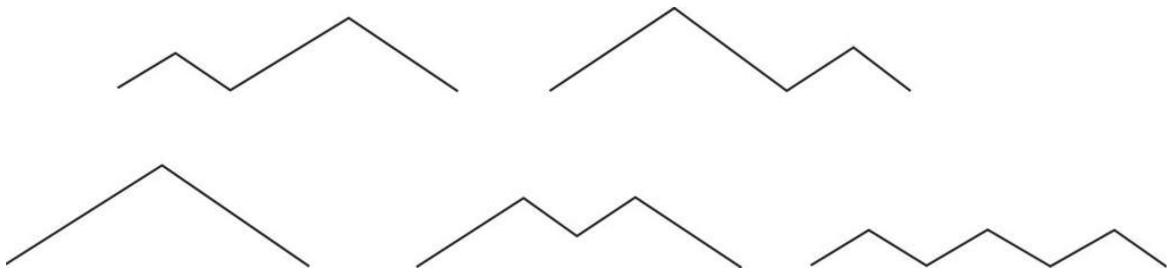
0, 1, 2, 3etc. fromasetof8. In total, this gives us the number of ways of selecting a set of any size from a group of 8, which is equal to 28 as, in general, a set of size ncontains 2n subsets within it.

This last fact can be seen directly, for a subset of a set of size ncan be identified by a binary string of length nin the following way. We consider the set in question in a specific order{ a1, a2, • • •, an} say, and then a binary string of length n specifies a subset by saying that each instance of 1 in the string indicates the presence of the corresponding a_i in the subset in question. For example, if n=4, the strings 0111 and 0000 stand respectively for{ a2, a3, a4}, and for the empty set. Since there are two choices for each entry in the binary string, there are 2^n such strings in all and therefore 2n subsets within a set of size n.

Catalannumbers

Every second row in the Arithmetic Triangle has a number sitting in the middle: 1, 2, 6, 20, 70, 252, 924, • • •. These numbers are divisible by the consecutive counting numbers 1, 2, 3, 4, 5, 6, 7, • • • and the numbers that come about as we carry out these divisions, 1, 1, 2, 5, 14, 42, 132 • • • are known as the Catalan numbers. In terms of these central binomial coefficients, thenth Catalan number can be expressed as $\frac{1}{n+1}C$ (2n, n) forn=0, 1, 2, • • •.

One of the simplest visual representations that gives rise to this number type is as the number of ways we can draw ‘mountains’ using nup strokes and n down strokes(see Figure 3)



3. With three up and down strokes there arefive mountain patterns

Each mountain pattern has an interpretation, however, as a meaningful bracketing and so the number of meaningful ways of arranging a collection of npairs of parentheses is the nth Catalan number. For example,(())()and((()))are meaningful bracketings but()))(()is not:to be meaningful, the number of left brackets must never fall behind the number of right brackets as we count from left to right. This corresponds to the natural condition that our mountains must never dive underground. For instance, the first and last mountain patterns in Figure 3 correspond to the bracketing()(())and()()()respectively.

The nth Catalan number also counts the number of ways that we can break up a regular polygon with n+2 sides into triangles by means of diagonals that do not cross one another, and there are other interpretations along these lines. As with binomial coefficients, there are formulas

relating Catalan numbers to smaller Catalan numbers, which makes them amenable to manipulation.

Fibonacci numbers

The Fibonacci sequence is one series of numbers that engenders wide fascination among the general public. The sequence runs as follows

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, \cdots$$

where each number after the pair of initial 1s is the sum of the two that come before. In this, there is a similarity with the binomial coefficients in that each term is the sum of two previous ones in the sequence, but the method of formation of the Fibonacci numbers is simpler:

$f_n=f_{n-1}+f_{n-2}$where f_n denotes the nth Fibonacci number and wefix $f_1=f_2=1$.We call such a formula that defines each member of a sequence in terms of its predecessors a recursion or a recurrence relation.

How does this sequence arise? It wasfirst introduced in 1202 by Leonardo of Pisa, better known as Fibonacci, in the form of his celebrated Rabbit Problem. A female rabbit is born and after two months reaches maturity and thereafter gives birth to a daughter each month. The number of female rabbits we have at the beginning of each month is then given by the Fibonacci numbers,for there is 1 rabbit at the beginning of thefirst month, and the second, but at the start of the third month she gives birth to a daughter so we then

have 2 rabbits. Next month she has another, giving 3 and the month after that we have 5 bunnies as both mother and her eldest daughter are now old enough to breed. In general, at the beginning of each month thereafter, the number of newborn daughters equals the number of females we had two months ago, as only they are old enough to breed. It follows that the number of females we have at the start of each subsequent month equals the total of the previous month(Fibonacci’s rabbits are immortal)plus the number we had the month before that.Therefore the rule of formation of the Fibonacci numbers exactly matches the breeding pattern of his rabbits.

Despite the fact that real rabbits do not breed in this contrived fashion, Fibonacci numbers arise in nature in a variety of ways, including plant growth. The reasons for this are well understood but are related to more subtle attributes of the sequence connected to the so-called Golden Ratio, a number that we are about to introduce.

The simplest types of number progressions are the arithmetic and geometric progressions introduced in thefirst section. Although the Fibonacci sequence is neither of these, it does however have a surprising link with the latter type. If we form the sequence of differences of the Fibonacci sequence, because of the way the sequence is defined, we get $0, 1, 1, 2, 3, 5, 8, 13, \cdots$, thatiswe recover the Fibonacci sequence again except this time beginning at 0. This happens precisely because of the way the sequence is formed: the difference of two consecutive Fibonacci numbers is the one

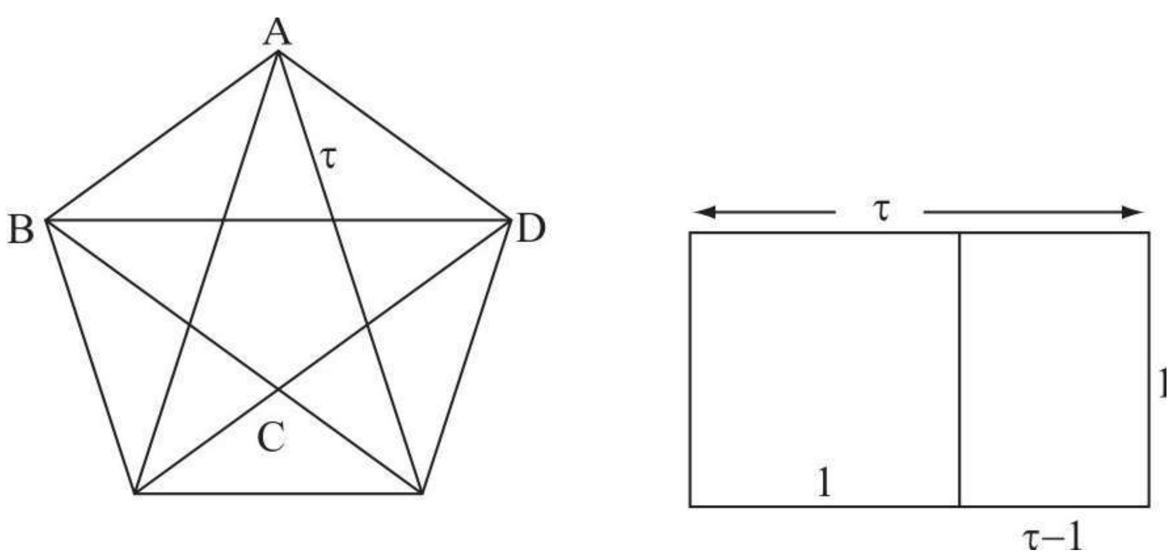
immediately preceding both in the sequence.(To see this algebraically, subtract f_{n-1} from both sides of the Fibonacci recurrence above.)Nor is the sequence a geometric progression as the ratio of consecutive Fibonacci numbers is not constant. All the same, when we look at the ratio of successive terms we see that it does seem to settle down to a limiting value. This near stable behaviour of the ratio comes about quite quickly, as we see as we divide each Fibonacci number by its predecessor:

$$\frac{34}{21} = 1 \cdot 6190,\ \frac{55}{34} = 1 \cdot 6176,\ \frac{89}{55} = 1 \cdot 6182,\ \frac{144}{89} = 1 \cdot 6180, \cdots$$

But what is the mysterious number, 1.6180..., whichwesee emerging?This numberτ is known as the Golden Ratio, andit arises quite of its own accord in geometrical settings that look a world away from Fibonacci's rabbits. For example, τ is the ratio of the diagonal of a regular pentagon to its side(see Figure 4). Each diagonal meets another at a point that divides each into two parts that are themselves in the ratioτ : 1. Pairs of intersecting sides and intersecting diagonals form the four sides of a rhombus(a 'square' parallelogram)ABCD as shown. Where diagonals cross, they form a smaller inverted pentagon.

A characteristic associated with the Golden Ratio isself-similarity, which means that objects such as the pentagon that involveτ often contain smaller copies of themselves within.This is seen in the rectangle with sides of

lengths τ and 1, for it is unique in displaying the property that if we slice off the largest square we can (a square of side length 1) then the smaller rectangle that remains is a copy of the original. This figure is for that reason known as the Golden Rectangle (see Figure 4). The value of τ can be gleaned from the given property of the rectangle, for if we call the length of the longer side τ and make the shorter side one unit, the similarity of the rectangles is captured by the equality $\frac{\tau}{1} = \frac{1}{\tau - 1}$, which we obtain by equating the ratio of the longer to the shorter side of the rectangles. Cross-multiplying in this expression then yields the equation $\tau^2 - \tau = 1$. Using the standard formula to solve such a quadratic equation (one involving a square), we see that the positive root of this equation equals:



4. Pentagonand the Golden Rectangle

$$\tau = \frac{1+\sqrt{5}}{2} = 1 \cdot 6180229 \cdots$$

Another way of retrieving this valuation of τ is through its so-called continued fraction, which ties τ directly to the Fibonacci numbers, and we shall explore this idea in Chapter 7.

In the long run, the Fibonacci sequence behaves like a geometric progression based on the Golden Ratio. It is this property, together with its simple rule of formation that causes the Fibonacci sequence to arise so persistently.

Stirling and Bell numbers

Like the binomial coefficients, the Stirling numbers often arise in counting problems and depend on two variables, nand r. The Stirling number S(n, r) is the number of ways of partitioning a set of nmembers into r blocks(with no block empty, and the order of the blocks and within the blocks, is immaterial). (Strictly these are called Stirling numbers of the second kind. Those of thefirst kind, which are related, count something quite different, namely the number of ways we can permute nobjects into r cycles.) For instance, the set with members a, b, c can be partitioned into three blocks in just one way: {a}, {b}, {c}, into two blocks in three ways {a, b}, {c}; {a}, {b, c}, and {a, c}, {b}, and into a single block in one way only: {a, b, c}; it follows that S(3, 1)=1, S(3, 2)=3 and S(3, 3)=1. Since a set of nmembers can be partitioned in only

one way into either 1 block or into nblocks, we always have S(n,1)=1=S(n,n). If we draw up the triangle of Stirling numbers after the fashion of Pascal's Triangle, we arrive at the array of Figure 5, and we now explain how the triangle is generated.

Once again, the numbers satisfy a recurrence relation, meaning that each can be related to earlier ones in the array. Indeed, as with the binomial coefficients, each Stirling number can be obtained from the two above it, but it is not simply the sum. What is more, the row symmetry we saw in the Arithmetic Triangle that generates the binomial coefficients is not present in Stirling's Triangle. For example, S(5,2)=15 but S(5,4)=10. The rule of recurrence is simple enough, however. The entry 90, for example, is equal to 15+3× 25. This is indicative of the general situation: tofind a number in the body of the triangle, take the two immediately above it, and add thefirst to the second multiplied by the number of the position in the row you are at. (This time, unlike the Arithmetic Triangle, start your row count at 1.) In a similar way, the entry S(5,4)=10=6+4× 1. It is only the part of the rule in italics that differs from that of the Arithmetic Triangle. That is enough, however, to make the study of Stirling numbers considerably more difficult to that of the binomial coefficients. For instance, we derived a simple explicit formula for each binomial coefficient in terms of the factorials. Similarly, there is a formula for the nth Fibonacci number in terms of powers of the Golden Ratio, but nothing of the kind exists for Stirling numbers.

1

1 1

1 3 1

1 7 6 1

1 15 25 10 1

1 31 90 65 15 1

1 63 301 350 140 21 1

⋮

5. Stirling’s Triangle

The recurrence rule is not hard to explain. We argue similarly to that for the recursion for the binomial coefficients, and by doing so recover the recurrence outlined above that is identical in form except for a single multiplier. In order to form a partition of a set of size ninto r non-empty blocks, we may proceed in two distinct ways. We may take thefirst n-1 elements of the set and partition it into r-1 non-empty blocks in S(n-1, r-1)ways, and thefinal member of the set will then form the rth block. Alternatively, we may partition thefirst n-1 elements of the set into r non-empty blocks, which can be done in S(n-1, r)ways, and then decide in which of the r blocks to place thefinal member of the set, giving a multiplier of r to that number. Hence we infer that

$$S(n,r)=S(n-1,r-1)+rS(n-1,r) \text{ for } n=2,3,\cdots$$

Using this recursion formula, we may calculate each line of the Stirling Triangle from the one above it. For example, putting n=7and r=5 we obtain:

$$S(7,5)=S(6,4)+5S(6,5)=65+5\times 15=65+75=140.$$

We can compute $S(n,2)$ and $S(n,n-1)$ directly from the definition as follows. An arbitrary partition of the n-set into afirst set and a second set is described by a binary string of length n, where the presence of a 1 indicates presence in thefirst set and a 0in the second(in a similar way to how we showed that the number of subsets of an n-set is 2^n). There are therefore 2^n such ordered pairs of sets. Since, however, there is no ordering of the blocks within a partition, we divide this number by 2 tofind the number of partitions of the n-set into 2 sets, giving the number 2^{n-1}. Finally, we need to subtract 1 from this in order to exclude the case where one of the sets is empty; hence $S(n,2)=2^{n-1}-1$. You can check that this represents the second diagonal line of numbers1, 3, 7, 15, 31, 63, • • • running from the top right to the bottom left in Figure 5.

At the other extreme, a partition of the n-set into n-1 blocks is determined by a choice of the unique block of size 2. The number of ways of making this selection is $C(n,2)=\frac{1}{2}n(n-1)$, the(n-1)st triangular number(see the second diagonal1, 3, 6, 10, 15, 21, • • • running from top left to bottom right in Figure 5).

The sum of any row of the Arithmetic Triangle gives the corresponding power of 2- the number of subsets of a set of a given size. Similarly, summing the nth row of Stirling' s Triangle gives the number of ways of breaking a set of nobjects into blocks, and this is called the nth Bell number.

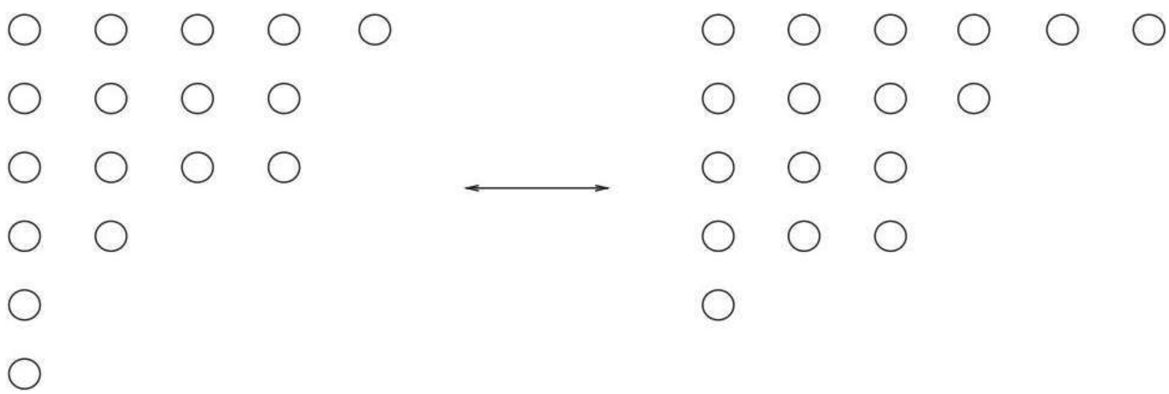
Partition numbers

If on the other hand, the nobjects of the set to be partitioned are identical, and so cannot be distinguished from one another, the number of ways of splitting the whole collection up into blocks is a much smaller integer, known as the nth partition number. A particular partition corresponds to writing nas a sum of positive integers, without regard to order: for example, 1+1+1+1+1 is one partition of 5 and there are six others, for we can also represent5as1+1+1+2, 1+2+2, 1+1+3, 2+3, 1+4, orsimply as 5. Therefore the 5th partition number is 7(that compares to the5th Bell number, which from the Stirling Triangle is seen to be1+15+25+10+1=52). There is no simple exact formula for the nth partition number- there is a complex one, which is itself based on a beautiful approximation due to the Indian genius Srinivasa Ramanujan(1887- 1920).

One simple symmetry regarding partitions is that the number of partitions of ninto mparts is equal to the number of partitions of nin which the largest part is m. One way of seeing that this is true is through the Ferrar' s graph(or Young diagram)of the partition, which is no more than the representation of the partition as a corresponding array of dots in which the rows are listed by decreasing size.

In the example shown in Figure 6 we have represented 17partitioned as 5+4+4+2+1+1. Note how the columns are also listed in decreasing order from left to right. If we reflect

the array along the diagonal running from top left to bottom right, we recover a second Ferrar' s graph as shown, which can be interpretedasthepartition17=6+4+3+3+1.Asimilar reflection of the second graph returns you to thefirst and we say that the two corresponding partitions are dual to one another.This symmetry allows us to see that the numbers of partitions of two corresponding types are equal:the dual of a partition in which m, say, is the largest number(so the top row has mdots)is a partition with mrows, which corresponds to a partition into m numbers. For example, the number of partitions of 17 into 6numbers therefore equals the number of partitions of 17 in which6 is the largest number that occurs.



6. Dualpartitionsof17=5+4+4+2+1+1=6+4+3+3+1

Mathematicians always have an eye out for these kind of symmetries that often arise in enumeration problems. Another example of this type occurs in relation to the Bertrand-Whitworth ballot problem where the votes for two electoral candidates are counted with the winner taking p votes and the loser q votes, say.A clever geometric argument using what is

known as the Reflection Principle shows that the proportion of counts where the winner leads the counting throughout the night equals the winner's final margin of victory divided by the total number of votes cast: $\frac{p-q}{p+q}$. This in turn is equal to the proportion of counts where the eventual winning margin is never attained until the very last vote is counted. The reason why these two numbers must be equal is that the two types of count are dual to one another in that the reversal of the order of votes in a count of the first type gives a count of the second type, and vice versa.

Hailstone numbers

Although not a counting tool, the hailstone numbers are intriguing as they are also defined recursively but have more of a flavour of the aliquot sequences that we met in Chapter 3. The following question goes by several names, the Collatz Algorithm, the Syracuse Problem, or sometimes just the 3n+1-problem, and it is simply the observation that, beginning with any number n, the following process always seems to end with the number 1. If n is even, divide it by 2, while if n is odd, replace it by 3n+1. For example, beginning with n=7 we are led by the rules through the following sequence:

$$7 \to 22 \to 11 \to 34 \to 17 \to 52 \to 26 \to 13 \to 40 \to 20 \to 10 \to 5 \to 16 \to 8 \to 4 \to 2 \to 1$$

And so the conjecture is true for n=7, and indeed it has been verified for all n up beyond a million million. Things are different if you fiddle with the rules: for instance, replacing 3n+1 by 3n−1 results in a cycle:

$$7 \to 20 \to 10 \to 5 \to 14 \to 7 \to \cdots.$$

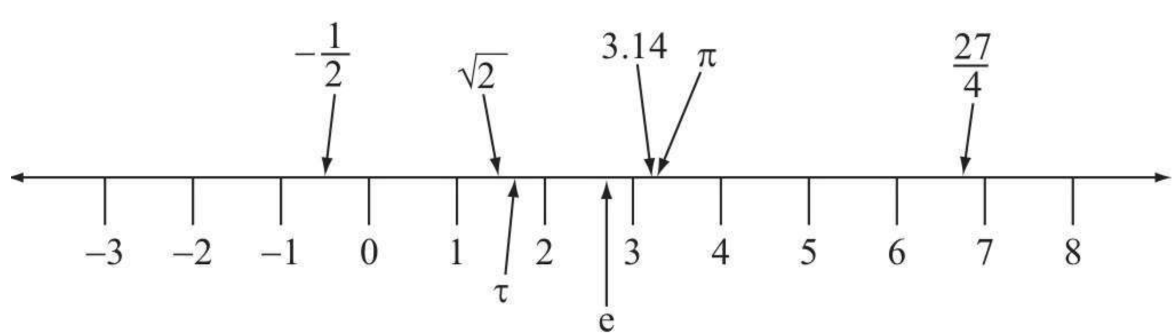
The sequences of numbers that arise from these calculations behave like hailstones in that they rise and fall erratically over a long period but eventually, it seems, always hit the ground. Of the first 1000 integers, more than 350 have a hailstone maximum height of 9232 before collapsing to 1. This will happen once you run into a power of 2, for they are exactly the numbers that cause you to fall straight down to ground level without encountering any more updrafts.

All sorts of intriguing features can be discerned in graphs and plots based on the hailstone sequences reminiscent of other chaotic patterns that arise in maths and physics. Typing 'hailstone numbers' into your favourite search engine will provide you with a wealth of information, often intriguing, sometimes speculative, but generally inconclusive.

Chapter 6 Below the waterline of the numbericeberg

Introduction

The counting numbers, 1, 2, 3, • • • are just the tip of the number iceberg. This tip is of course thefirst part we discover, and for a time we might believe there is no more to the iceberg than the tip, especially if we remain reluctant to look below the waterline. In the course of this chapter, wefirst introduce the negative integers and coupling this extension with fractions, both positive and negative, gives the collection we call the set of rational numbers. This number collection is often pictured to lie along the numberline, with the positive numbers lying to the right of zero, with their negatives forming the mirror image to the left. However, the number line turns out to be the home of other numbers that cannot be expressed as fractions, such as $\sqrt{2}$ and π , totaketwo examples. The set of real numbers is the name for the collection of all numbers on the number line, which are those that can be represented by decimal expansions of any kind, as shown in Figure 7.



7. Central portion of the number linenear 0

However, one of the great achievements of the 19th century was the full realization that the true domain of number is not one-, but rather is two-dimensional. The plane of the complex numbers is the natural arena of discourse for much of mathematics. This has been brought home to mathematicians and scientists through problem solving- to be able to carry out the investigations required to solve real-world problems, many of which seem to be only about ordinary counting numbers, it becomes necessary to expand your number horizon. The explanation as to how this extra dimension emerges will come towards the end of this chapter and be explored further in Chapter 8.

Pluses and minuses

The integers is the name applied to the set of all whole numbers,positive negative, and zero. This set, often symbolized by the letter Z, is therefore infinite in both directions:

$$\{\cdots -4, -3, -2, -1, 0, 1, 2, 3, 4, \cdots\}.$$

The integers are often pictured as lying at equally spaced points along a horizontal number line, in the order indicated. The additional rules that we need to know in order to do arithmetic with the integers can be summarized as follows:

(a) to add or subtract a negative integer, −m, we move m spaces to the left in the case of addition, and m spaces to the right for subtraction;

(b) to multiply an integer by −m, we multiply the integer by m, and then change sign.

In other words, the direction of addition and subtraction of negative numbers is the opposite to that of positive numbers, while multiplying a number by −1 swaps its sign for the alternative. For example, 8+(−11)=−3, 3×(−8)=−24, and (−1)×(−1)=1.

You should not be troubled by this last sum. First, it is reasonable that multiplying a negative number by a positive one yields a negative answer: when a debt (a negative amount) is subject to interest (a positive multiplier greater than 1) the outcome is greater debt, that is to say a larger negative number. We are all well aware of this. That multiplication of a negative number by another negative number should have the opposite outcome, that is a positive result, would then appear consistent. The fact that the product of two negative numbers is positive can readily be given formal proof. The proof is based on the assumptions that we want our expanded number system of the integers to subsume the original one of the natural numbers, and that the augmented system should continue to obey all the normal rules of algebra. Indeed, the result on the product of two negatives follows from the fact that any number multiplied by zero equals zero. (This too is

not an assumption but rather is also a consequence of the laws of algebra.)For we now have:

$$-1\times(-1+1)=-1\times 0=0;$$

if we then multiply out the brackets, we see that in order that the left-hand side equal zero, $(-1)\times(-1)$ must take the opposite sign to $(-1)\times 1=-1$; in other words $(-1)\times(-1)=1$.

Fractions and rationals

In a similar way that subtraction leads to the negative numbers, the operation of division also leads us out of the set of natural counting numbers into the larger realm of fractions. However, the nature of the new arithmetic we encounter is of a different character. When adding or subtracting, fractions with different denominators(bottom lines)are incompatible. The fractions in question need to be expressed with a common denominator before the sum can be completed. Multiplication is a comparatively simple process in that we only need to multiply the numerators(top lines)and denominators together in order to get the answer. Division is the inverse operation to multiplication so that division by n corresponds to multiplication by the reciprocal, $\frac{1}{n}$. In general, this carries over to fractions in that to divide by the fraction $\frac{m}{n}$ we multiply by its reciprocal, $\frac{n}{m}$ for that reverses the effect of multiplication by $\frac{m}{n}$

The Ancient Egyptians were only happy with unit fractions, which are those that are simple reciprocals of

whole numbers, $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$ etc. (although they retained a special symbol for $\frac{2}{3}$). A fraction such as $\frac{3}{4}$ was not thought of as a meaningful entity in its own right, and they would record this quantity as the sum of two reciprocals: $\frac{3}{4} = \frac{1}{2} + \frac{1}{4}$. (The notation for fractions used here is of course the modern European type, which has its origins in Greek mathematics.) It is not, however, obvious that it is necessarily possible to write any fraction as the sum of a number of different unit fractions, which is what they insisted on. It can, however, always be done and explaining this will allow you to brush up your skills on dealing with fractions.

If you wish tofind an Egyptian decomposition of a fraction such as $\frac{9}{20}$, you need only subtract the largest unit fraction you can from the given number, and repeat this process until the remainder is itself a unit fraction. This will always work, and the number of fractions involved never exceeds the numerator of your original fraction. This is because, at each stage, the numerator of the fraction that still remains is always less than the previous one: not obvious but true. In this example thefirst stage will give:

$$\frac{9}{20} - \frac{1}{3} = \frac{27}{60} - \frac{20}{60} = \frac{7}{60};$$

next wefind that the largest unit fraction less than $\frac{7}{60}$ is $\frac{1}{9}$. (Totest this, compare the result of the cross-multiplication: $\frac{1}{9} < \frac{7}{60}$ because 1× 60=60＜63=7× 9.) Subtracting again, we see that

$$\frac{9}{20} - \frac{1}{3} - \frac{1}{9} = \frac{7}{60} - \frac{1}{9} = \frac{21}{180} - \frac{20}{180} = \frac{1}{180}$$

and so we recover the Egyptian decomposition:

$$\frac{9}{20} = \frac{1}{3} + \frac{1}{9} + \frac{1}{180}.$$

This greedy approach of always subtracting the largest unit fraction available does work but may not yield the shortest decomposition there is, as we can see even in this case as $\frac{9}{20} = \frac{1}{4} + \frac{1}{5}$. This two-fraction decomposition of $\frac{9}{20}$ can be found, however, through use of the technique of the Akhmim papyrus, a Greek parchment discovered at the city of Akhmim on the Nile and dated to AD 500- 800. In modern notation, the trick can be expressed as the readily verified algebraic identity:

$$\frac{m}{pq} = \frac{m}{p(p+q)} + \frac{m}{q(p+q)}.$$

Applying this with m=9, p=4, q=5 immediately gives us

$$\frac{9}{20} = \frac{9}{4\times 9} + \frac{9}{5\times 9} = \frac{1}{4} + \frac{1}{5}.$$

It may well have been years since you last added even simple fractions together because nearly all practical arithmetic is now carried out in decimal format. The use of decimal fractions is found in ancient China and medieval Arabic nations but only came into widespread use in Europe in the latter part of the 16th century when serious efforts were made to improve practical methods of computation. There is a price to be paid, however, for this commitment to decimal forms. In normal base ten arithmetic we exploit the fact that any number can be written as a sum of multiples of powers of ten. When expressing a fraction as a decimal, we are attempting to write the number as a sum of powers of $\frac{1}{10}=0.1$. Unfortunately, even for very simple fractions such as $\frac{1}{3}$, this cannot be done, and the decimal expansion goes on without end: $\frac{1}{3}=0.333...$. In practice, we appreciate that by truncating the decimal expansion after a certain number of places(depending on the accuracy we demand), we can get by with the resulting terminating decimal that approximates the exact fraction. Any inaccuracy is trivial in comparison with the convenience of carrying out all our number work in the standard base ten frame of reference. Decimal expansions can be thought of as the closest we can get to having a single common denominator for all fractions.

It is natural to ask, though, which fractions will have terminating expansions(and which will not)?The answer is, not very many.More often than not, the decimal expansion of a fraction goes into a recurring pattern: $\frac{3}{22}=0.1363636...$ with

the 36 part repeating forever. Every fraction generates a recurring decimal in one way or another, although in the case of a terminating decimal such as $\frac{1}{2}$=0.5, the recurring part is simply an unending string of zeros: $\frac{1}{2}$=0.5000···, and so is not explicitly mentioned. In any event, the length of the recurring block in a recurring decimal expansion is no longer than one less than the value of the denominator. This can be seen by considering what happens when we carry out the corresponding long division sum: if the denominator is n, then the remainder after each step in the division takes on one of the values 0, 1, ···, n−1. If at some stage the remainder is 0, the division terminates and so does the decimal expansion: for example, $\frac{11}{40}$ is exactly equal to 0.275. Otherwise the division continues forever, but once a remainder is repeated, which is inevitable, we shall be forced into the same cycle of divisions once more, thus giving a recurring pattern whose block can be no longer than n−1. The expansion will terminate exactly when the denominator is a product of the prime factors 2 and 5 of our base 10 but not if there is any other factor involved. For example, fractions with denominators 16, 40, and 50 are terminating, but fractions like $\frac{1}{14}$ and $\frac{1}{15}$ will not terminate because the respective prime factors of 7 and 3 in their denominators force the expansion into a recurring cycle.

This does show, however, that whether or not a fraction's expansion terminates is not determined just by the number itself but rather depends on the relationship of

the number to the base in which you are operating. If, for instance, we worked in ternary(base three)then 0.1 would represent $\frac{1}{3}$, as the 1 after the decimal point would stand for $\frac{1}{3}$, and not $\frac{1}{10}$, thewayitdoesindecimal expansions.

The reverse process of turning a recurring decimal back into a fraction is also quite simple, showing that there is a one-to-one correspondence between fractions and recurring decimals, and we can use whichever representation best suits our current purpose. A simple example is as follows: let a=0.212121•••. Sincethe length of the recurring block is 2, we can simplify this, as you will see, by multiplying by 10^2=100 to obtain 100a=21.212121•••. This has been set up so that, upon subtraction, the recurring parts of the two numbers a and 100a, being identical, will cancel, allowing us to infer that 99a=21, whence a= $\frac{21}{99} = \frac{7}{33}$.

This kind of trick is often used to simplify an expression that involves an infinite repeating process. For example, consider the following little monster:

$$a = \sqrt{2\sqrt{5\sqrt{2\sqrt{5}\cdots}}}$$

By squaring, and then squaring again, the left-hand side becomes a4, while the expression on the right gives:

$$a^4 = 2^2 \times 5 \times \sqrt{2\sqrt{5\sqrt{2\sqrt{5}}}} \cdots$$

Since what follows the 5 is another copy of the expression for a, we infer that a4=20a so that a3=20 or, if you prefer, a is the cube root of 20. We will call on this technique again in Chapter 7 when we introduce so-called continued fractions.

Does the class of fractions provide us with all the numbers we could ever need?As mentioned earlier, the collection of all fractions, together with their negatives, form the set of numbers known as the rationals, that is all numbers that result from whole numbers and the ratios between them. They are adequate for arithmetic in that any sum involving the four basic arithmetic operations of addition, subtraction, multiplication, and division will never take you outside the world of rational numbers. If we are happy with that, this set of numbers is all we require. However, we explain in the next section how numbers such as a above are not rational.

Irrationals

The word irrational applied to a number a means simply that the number is not rational, that is to say cannot be written as a fraction. Irrational numbers werefirst discovered a very long time ago, in ancient Greece. Pythagoras understood the irrational nature of $\sqrt{2}$ The Greeks did not think in terms of decimal expansions but were happy to recognize a length constructed in the geometry of straightedge and compass as representing a real quantity. In particular, Pythagoras' Theorem tells us that the longer side of a right-angled triangle whose shorter sides are each of length 1 unit is exactly equal to the square root of 2.

Pythagoras was able to prove that the square root of 2 was not equal to any fraction, thereby showing that irrational numbers truly exist. In particular, you cannot exactly measure the diagonal of a square with the same units with which you measure the side, for if you could the diagonal would be an exact fractional multiple of the side, in which case $\sqrt{2}$ would be equal to this fraction. The two lengths are however fundamentally incompatible, orincommensurable as they are described in the classical texts. The story is the same for π, which is approximately equal to the fraction $\frac{22}{7}$, but is different from it, and from any fraction that you care to nominate.(However, the easily remembered 'double 1, double 3, double 5' ratio: 355/113=3.1415929・・・accurately approximates the value of π to better than one part in one million.)Although it is very difficult to prove that π is irrational, the question for the square root of 2 can be settled easily by a simple contradiction argument. First, we

note that for any number c, the highest power of 2 that is a factor of c^2, is twice the highest power of 2 that is a factor of c, and so in particular the highest power of 2that divides any square must itself be an even number. For example, $24=2^3\times 3$while$576=24^2=2^6\times 3^2$, and in this case the highest power of 2 dividing the number does indeed double from 3 to 6 when we take the square. This is always the case, and indeed applies not only to powers of 2 but to any prime factor of the original number.

Suppose now that $\frac{a}{b}$were equal to the fraction $\frac{a}{b}$squaring both sides of this equation allows us to deduce that $2=\frac{a^2}{b^2}$, which gives$2b^2=a^2$. By the previous observation, the highest power of 2 that divides the right-hand side of this equation is even, while the highest power that divides the left-hand side is odd(because of the presence of the extra 2). This shows the equation to be nonsense, and so it must not be possible to write $\sqrt{2}$as a fraction in thefirst place. Like Pythagoras, we come face to face with the irrationals.

Arguments along these lines allow us to show that quite generally, when we take the square root(or indeed the cube or a higher root)of a number, the answer, if not a whole number, is always irrational, thus explaining why the decimal displays on your calculator never show a recurring pattern when asked to calculate such a root.

Pythagoras discovered that in order to do his mathematics, he required a widerfield of numbers than mere fractions. The Greeks regarded a number to be ‘real’ if its length could be constructed from a standard unit interval using only a straightedge(not a marked ruler, just an edge)and compass. It turns out that although the square root operation does introduce irrationals, the full collection itself does not go very far beyond the rational. The set of euclidean numbers, as we shall refer to them, are all those that can be arrived at from the number 1 through carrying out any or all of the four operations of arithmetic and the taking of square roots, any number of times. For example, the number $\sqrt{7-\sqrt{4/3}}$ is therefore a number of this kind. Even cube roots are beyond the grasp of the euclidean tools. This was the basis of perhaps the first great unsolved problem in mathematics. Thefirst of the three Delian Problems as they were known was the call to construct the cube root of 2, using only straightedge and compass. Legend has it that this was the task set by the god when the citizens of Delos consulted the Oracle of Delphi to learn what they should do in ordertobanishtheplaguefromAthens- theproblemwasputin the form of exactly doubling the volume of an altar that was a perfect cube.

This problem remained untouchable in classical times. That the cube root of 2 lies outside the range of the euclidean tools was only settled in 1837 by Pierre Wantzel(1814- 38), as it requires a precise algebraic description of what is possible using the classical tools in

order to see that the cube root of 2 is a number of a fundamentally different type. It does indeed come down to showing that you can never manufacture a cube root out of square roots and rationals. When put that way, the impossibility sounds more plausible. However, that in no way constitutes a proof.

Transcendentals

Within the class of irrationals lies the mysterious family of transcendental numbers. These numbers do not arise through the ordinary calculations of arithmetic and the extraction of roots. For the precise definition, wefirst introduce the complementary collection of algebraic numbers, which are those that solve some polynomial equation with integer coefficients: for example $x^5-3x+1=0$ is such an equation. The transcendentals are then defined to be the class of non-algebraic numbers.

It is not at all clear that there are any such numbers. However, they do exist and they form a very secretive society, with those in it not readily divulging their membership of the club. For example, the number π is an instance of a transcendental but this is not a fact that it openly reveals. It will be explained in the next chapter when we explore the nature of infinite sets why it is that 'most' numbers are transcendental, in a sense that will be made precise.

For the time being, I will settle for introducing perhaps the most famous transcendental of all, the number

e=2.71828....This number arises constantly in higher mathematics and calculus:it is the base of the so-called natural logarithm, the function that tells you the area under the graph of the reciprocal function. It is also the limiting value of the sequence of numbers you get when you raise the ratio of two consecutive integers, $\frac{n+1}{n}\ (=1+\frac{1}{n})$, to the power n.(Ask your calculator for the value of $(129/128)^{128}$- you can 'fast exponentiate' this, just calculate 129/128 and then square7 times, as 27=128.)

This sequence arises when we consider the problem of the limiting value of a compound interest rate as you reduce the interval of repayment shorter and shorter from annually, to monthly, to daily,and so on. To best illustrate the point, suppose that interest is paid at an annual rate of 100%, compounding in ninstalments per year,which means that your initial investment is multiplied by the factor $(1+\frac{1}{n})$, ntimes in all, throughout the course of the year.Your principal will then be multiplied by the factor $(1+\frac{1}{n})^n$. .The more often you are paid interest, the more you will earn as you begin to collect interest on your interest earlier and earlier as n becomes higher and higher. However, as nincreases, the effective APR(Annual Percentage Rate)does not increase beyond all bounds but rather approaches a ceiling, an upper limit as mathematicians call it. This limiting multiplier that would apply to your principal in the continuous interest case is the limiting value, as nincreases, of the number

$$\left(1+\frac{1}{n}\right)^n \to e = 2.71828\ldots.$$

Another way in which the mysterious e arises is through the sum of the reciprocals of the factorials, and this gives a way of calculating e to a high degree of accuracy as this series converges rapidly because its terms approach zero very quickly indeed:

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots$$

This representation allows you to show by a relatively simple contradiction argument, outlined here, that e is an irrational number. We suppose that the preceding series for e equals a fraction $\frac{p}{q}$ and then we multiply both sides by q!. The left-hand side is then an integer but the right-hand side consists of terms that are integers followed by an infinite sequence of non-integral terms. By comparing to a simple geometric series, we deduce that this 'tail' sums to less than 1, and so the right-hand side cannot be a whole number, and therein lies the required contradiction. Showing that e is not just irrational, but transcendental, requires quite a bit more work.

The relationship of e with the factorials also manifests itself in a remarkable formula of the Scottish mathematician James Stirling(1692- 1770), after whom Stirling numbers(see Chapter 5)are named. He showed that as n increases, the value of n! is approximated better and better by the expression $\sqrt{2\pi n}(\frac{n}{e})^n$.

Since e crops up in a variety of distinct and fairly simple ways, it persistently appears throughout mathematics, often where you would not expect to meet it. For example, take two well-shuffled packs of playing cards, turn over the top card of each deck, and compare. Continue doing this until you have exhausted the packs.What are the chances that, at some stage, there is a perfect match?That is to say, on one turn or another the cards showing are exactly the same, be they the seven of clubs, queen of hearts, or whatever.It works out that the proportion of times this experiment yields at least one such match is as near as makes no difference to $\frac{1}{e}$, which is about 36.8%. This comes about through application of what is known as the inclusion- exclusion principle, which arrives at the solution through a sum of terms each of which represent alternating corrections and reverse corrections. In this example,the principle furnishes the series of reciprocals of the factorials but this time with alternating signs, which converges to $\frac{1}{e}$

Therealandtheimaginary

Thefirstfive chapters of this *Very Short Introduction* dealt mainly with positive integers. We emphasized factorization properties of integers, which led us to consider numbers that have no proper factorizations, which are the primes, a set that occupies a pivotal position in modern cryptography. We also looked at particular types of numbers, such as the Mersenne primes, which are intimately connected with perfect numbers and took time to introduce some special classes of integers that are important in counting certain naturally occurring collections. Throughout all this, the backdrop was the system of integers, which are the counting numbers, positive, negative, and zero.

In this chapter we have gone beyond integers, first to the rationals(the fractions, positive and otherwise), then to the irrationals, and within the class of irrationals we have identified the transcendental numbers. The underlying system in which all this is taking place is the system of the real numbers, which can be thought of as the collection of all possible decimal expansions. Any positive real number can be represented in the form $r=n.a_1a_2\cdots$, where n is a non-negative integer and the decimal point is followed by an infinite trail of digits. If this trail eventually falls into a recurring pattern, then r is in fact rational and we have shown how to convert this representation into an ordinary fraction. If not, then r is irrational, so the real numbers come in those two distinct flavours, the rational and the irrational.

In our mathematical imaginations, we often picture the real numbers as corresponding to all the points along the number line as we look out from zero, to the right for the positive reals, and to the left for the negative reals. This leaves us with a symmetrical picture with the negative real numbers being a mirror image of the positive reals, and this symmetry is preserved when dealing with addition and subtraction- but not with multiplication. Once we pass to multiplication, the positive and negative numbers no longer have equal status as the number 1 is endowed with a property that no other number possesses, for it is themultiplicative identity, meaning that $1\times r=r\times 1=r$ for any real number r. Multiplication by 1fixes the position of any number, but in contrast multiplication by-1 swaps a number for its mirror image on the far side of 0. Once multiplication enters the scene, the fundamental differences in the nature of positive and negative numbers are revealed. In particular, negative numbers lack square roots within the real number system because the square of any real number is always greater than or equal to zero.

This is the cue for imaginary numbers to make their entrance. This topic is one that we shall take up again in thefinal chapter; for the time being, we will just make some introductory comments. These numbers arise through the search for solutions to simple polynomial equations. In particular, since the square of any real number is never negative, we canfind no solution to the equation $x^2=-1$. Undaunted, mathematicians invented one, denoted by i, which is endowed

with that property, so that $i^2=-1$. Atfirst sight, this seems artificial and arbitrary but it is not too much different from the kind of behaviour we have indulged in before. After all, while recognizing that the counting numbers 1, 2, 3, • • • are pre-eminent, in order to deal smoothly with general number matters we are led to the wider number system of the rationals, which is the collection of all fractions, positive, negative, and zero. We thenfind, however, that we have no solution to the equation $x^2=2$, as we have shown that the square of a rational number cannot exactly equal 2. To deal with this, we have to 'invent' $\sqrt{2}$. At this point we could take the alternative attitude and say that we have proved that the square root of 2 simply does not exist and that is the end of the matter. However, few would feel happy to haul up the drawbridge in this way. The ancient Greeks certainly were not content to let things stand at that, for they could construct a length representing $\sqrt{2}$ with compass and straightedge and so the number was, to their way of thinking, definitely real and any mathematical system that denied this was inadequate.

We might agree with Pythagoras for quite a different reason. We may react by saying that we can approximate $\sqrt{2}$ toanydegreeof accuracy via its decimal expansion: $\sqrt{2}$ =1.414213 • • •, andso $\sqrt{2}$ is the number that is represented exactly by the totality of this expansion. A modern person mightfind more force in this argument and so insist for this reason that the number system needs to be expanded beyond the rationals.

However, atfirst glance we might say that things are different when it comes to $\sqrt{-1}$ as there seems no immediate need to worry about its non-appearance among the collection of numbers that we habitually call ‘real’. It transpires though that as our mathematics progresses a little further, the need for imaginary numbers becomes very pressing, and any initial reluctance to deal with them is dispelled as our understanding of things mathematical grows.

Thisfirst struck home in the 16th century when Italian mathematicians learnt how to solve cubic and fourth-degree polynomial equations in a fashion that extended that used to solve quadratic equations. The Cardano method, as it came to be known, would often involve square roots of negatives even though the solutions to the equations eventually turned out to be positive integers. By stages from this point, the use of complex numbers, which are those of the form a+bi, wherea and b are ordinary real numbers, was shown to facilitate a variety of mathematical calculations. For example, in the 18th century Euler revealed and exploited the stunning little equation $e^{i\pi}=-1$, which cannot fail to surprise anyone on theirfirst encounter.

Around the beginning of the 19th century, the geometric interpretation of complex numbers as points in the coordinate plane(the standard system of xy-coordinates), was investigated by Wessell and Argand, from which point the use of the ‘imaginary’ became accepted as normal mathematics. Identifying the complex number x+iywith the point with

coordinates(x, y)allows examination of the behaviour of complex numbers in terms of the behaviour of points in the plane, and this proves to be very illuminating. The theory of so-called complex variables, whose subject matter is represented by functions of complex numbers, rather than just real numbers, flourished spectacularly in the hands of Augustin Cauchy(1789- 1857). It is now a cornerstone of mathematics, underpins much of electrical signal theory, and the entirefield of X-ray diffraction is built on complex numbers.These numbers have proved to have real meaning, and moreover the system is complete in that every polynomial equation has its full complement of solutions within the system of complex numbers. We shall return to these matters in thefinal chapter.Before doing that, however, we shall in the next chapter look more closely at the infinite nature of the real number line.

Chapter 7 To infinity and beyond!

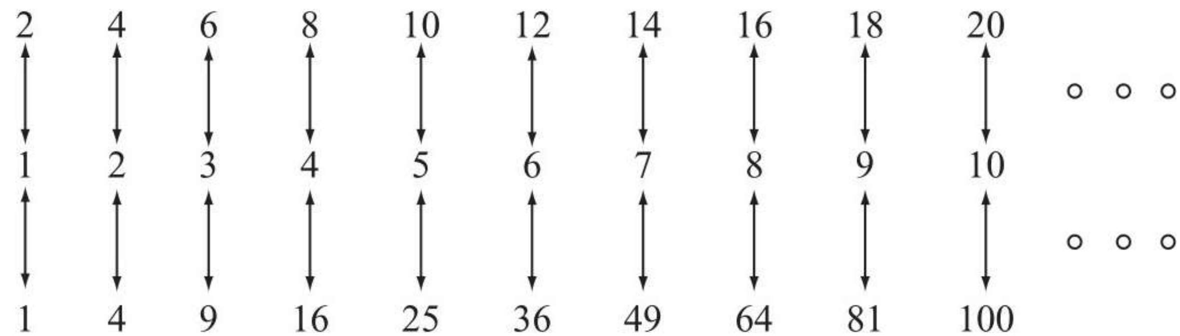
Infinity within infinity

It was the great 16th-century Italian polymath Galileo Galilei(1564- 1642)who wasfirst to alert us to the fact that the nature of infinite collections is fundamentally different fromfinite ones. As alluded to on thefirst page of this book, the size of afinite set is smaller than that of a second set if the members of thefirst can be paired off with those of just a portion of the second. However, infinite sets by contrast can be made to correspond in this way to subsets of themselves(where by the term subset I mean a set within the set itself). We need go no further than the sequence of natural counting numbers 1, 2, 3, 4, • • • in order to see this. It is easy to describe any number of subsets of this collection that themselves form an infinite list, and so are in a one-to-one correspondence with the full set(see Figure 8): the odd numbers, 1, 3, 5, 7, • • •, the square numbers, 1, 4, 9, 16, • • • and, less obviously, the prime numbers, 2, 3, 5, 7, • • •, and in each of these cases the respective complementary sets of the even numbers, the non-squares, and the composite numbers are also infinite.

The Hilbert Hotel

This rather extraordinary hotel, which is always associated with David Hilbert(1862- 1943), the leading German mathematician of his day, serves to bring to life the strange

nature of the infinite. Its chief feature is that it has infinitely many rooms, numbered1, 2, 3, • • •, and boasts that there is always room at Hilbert’ s Hotel.



8. The evensand thesquarespaired with the natural numbers

One night, however, it is in fact full, which is to say each and every room is occupied by a guest and much to the dismay of the desk clerk, one more customer fronts up demanding a room. An ugly scene is avoided when the manager intervenes and takes the clerk aside to explain how to deal with the situation: tell the occupant of Room 1 to move to Room 2 says he, that of Room 2 to move into Room 3, and so on. That is to say, we issue a global request that the customer in Room nshould shift into Room n+1, and this will leave Room 1 empty for this gentleman!

And so you see, there is always room at the Hilbert Hotel. But how much room?

The next evening, the clerk is confronted with a similar but more testing situation. This time a spaceship with 1729 passengers arrives, all demanding a room in the already fully

occupied hotel.The clerk has, however, learned his lesson from the previous night and sees how to extend the idea to cope with this additional group.He tells the person in Room 1 to go to Room 1730, that of Room 2to shift to Room 1731, and so on, issuing the global request that the customer in Room nshould move into Room number n+1729.This leaves Rooms 1 through to 1729 free for the new arrivals, and our clerk is rightly proud of himself for dealing with this new version of last night’s problem all by himself.

Thefinal night, however, the clerk again faces the same situation- a full hotel, but this time, to his horror, not just a few extra customers show up but an infinite space coach with infinitely many passengers, one for each of the counting numbers 1,2,3.•••.The overwhelmed clerk tells the coach driver that the hotel is full and there is no conceivable way of dealing with them all. He might be able to squeeze in one or two more, anyfinite number perhaps,but surely not infinitely many more. It is plainly impossible!

An infinite riot might have ensued except again for the timely intervention of the manager who, being well versed in Galileo’s lessons on infinite sets, informs the coach driver that there is no problem at all. There is always room at Hilbert’s Hotel for anyone and everyone. He takes his panicking desk clerk aside for another lesson. All we need do is this, he says. We tell the occupant of Room 1 to shift into Room 2, that in Room 2 to shift to Room 4,that in Room 3 to go to Room 6, and so on. In general the global instruction

is that the occupant of Room nshould move into Room 2n. This will leave all the odd numbered rooms empty for the passengers of the infinite space coach. No problem at all!

The manager seems to have it all under control. However, even he would be caught out if a spaceship turned up that somehow had the technology to have one passenger for each point in the continuum of the real line. One person for every decimal number would totally overrun Hilbert’ s Hotel, and we shall see why in the next section.

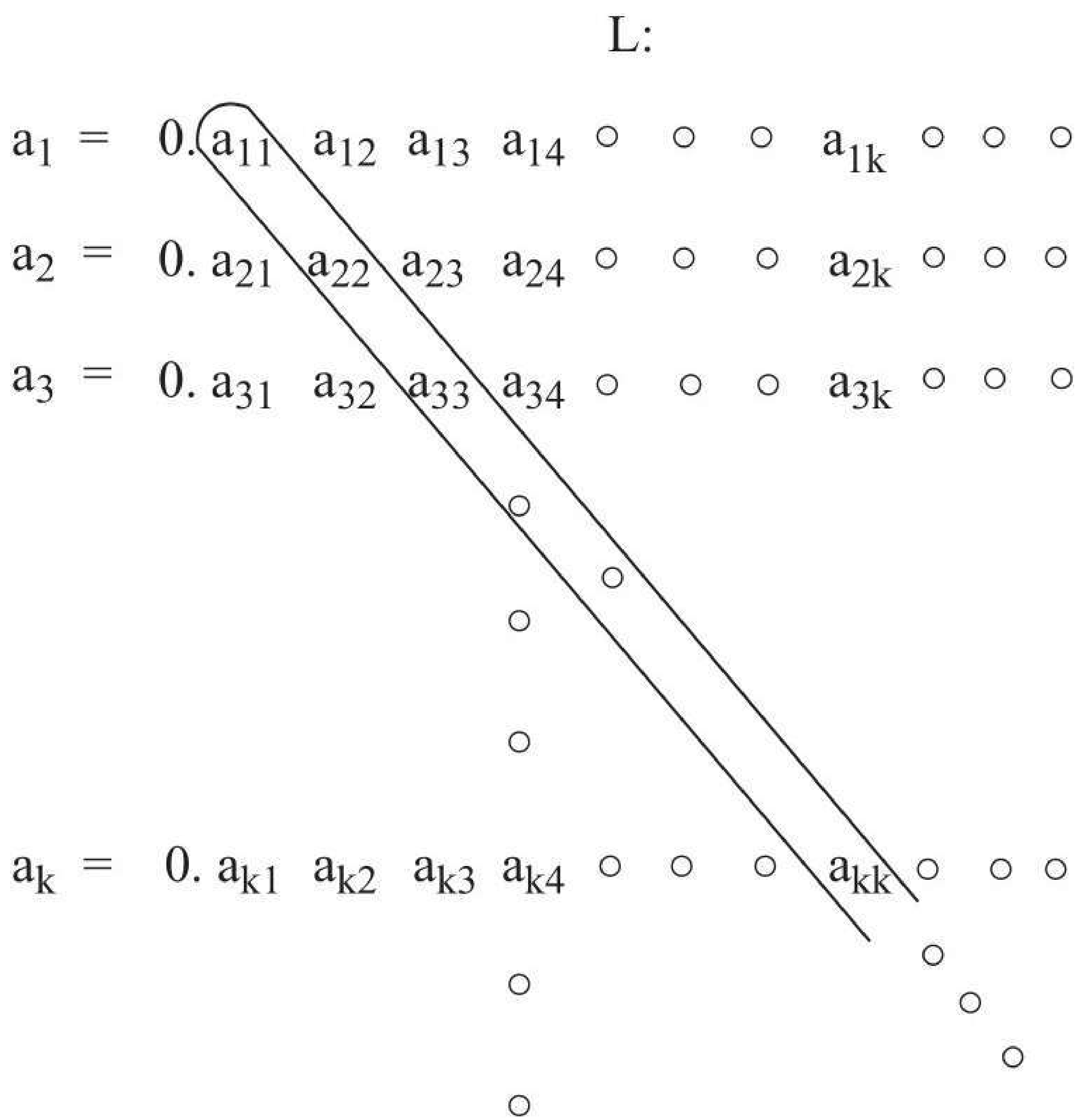
Cantor’ s comparisons

All this may be surprising thefirst time you think about it, but it is not difficult to accept that the behaviour of infinite sets may differ in some respects fromfinite ones, and this property of having the same size as one of its subsets is therefore a case in point. In the19th century, however, Georg Cantor(1845- 1918)went much further and discovered that not all infinite sets can be regarded as having equally many members. This revelation was unexpected in nature but is not hard to appreciate once your attention is drawn to it.

Cantor asks us to think about the following. Suppose we have any infinite list L of numbers a1, a2, • • • thought of as being given in decimal form. Then it is possible to write down another number, a, that does not appear anywhere in the list L: we simply take a to be different from a1 in thefirst place after the decimal point, different from a2 in the second

decimal place, different from a3 in the third decimal place, and so on- in this way, we may build our number a making sure it is not equal to any number in the list. This observation looks innocuous but it has the immediate consequence that it is absolutely impossible for the list L to contain all numbers, because the number a will be missing from L. It follows that the set of all real numbers, thatisalldecimal expansions, cannot be written in a list, or in other words cannot be put into a one-to-one correspondence with the natural counting numbers, the way we saw in Figure 8. This line of reasoning is known as Cantor’s Diagonal Argument, as the number a that lies outside the set L is constructed by imagining a list of the decimal displays of L as in Figure 9 and defining a in terms of the diagonal of the array.

There is some subtlety here, for we might suggest that we can easily get around this difficulty by simply placing the missing number a at the front of L. This creates an enlarged listing M containing the annoying number a. However, the underlying problem has not gone away. We can apply Cantor’s construction again to introduce a fresh number b that is not present in the new list M. We can of course continue to augment the current list as before any number of times, but Cantor’s point remains valid: although we can keep creating lists that contain additional numbers that were previously overlooked, there can never be one specific list that contains every real number.



9. Cantor’s numbera differs from each ak in the kth decimal place

The collection of all real numbers is therefore larger in a sense than the collection of all positive integers. Even though both are infinite, the sets cannot be paired off together the way the even numbers can be paired with the list of all counting numbers. Indeed, if we apply Cantor’s

Diagonal Argument to a putative list of all numbers in the interval 0 to 1, the missing number a will also lie in this range. Therefore, we likewise conclude that this collection will also defy every attempt at listing it in full. I mention this as we shall make use of that fact shortly.

Cantor's result is rendered all the more striking by the fact that many other sets of numbers can be put into an infinite list, including the Greeks' euclidean numbers. A little ingenuity is involved, but once a couple of tricks are learned, it is not hard to show that many sets of numbers are countable, whichistheterm we use to mean the set can be listed in the same fashion as the counting numbers. Otherwise an infinite set is called uncountable.

For example, let us take the set of all integers Z, which comes to us naturally as a kind of doubly infinite list. We can, however, rearrange it into a row with a starting point: 0, 1, −1, 2, −2, 3, −3, • • • by pairing each positive integer with its opposite, we create a list where every integer appears- none will escape. More surprisingly, we can also do the same with the rationals: start with0, then list all the rationals that can be written using all numbers no more than 1(which are 1 and−1), then those that involve no number higher than 2(which are 2, −2, $\frac{1}{2}, -\frac{1}{2}$), then those that only use numbers up to 3, and so on. In this way, the fractions(positive, negative, and zero)can be arranged in a sequence in which they are all present and accounted for. The rationals therefore also form a countable set, as do the euclidean numbers, and indeed if we consider the set of all numbers that arise from the rationals through taking roots of any order, the collection produced is still countable. We can even go beyond this: the collection of all algebraic numbers(first mentioned in Chapter 6), which are those that are solutions of ordinary polynomial equations, form a collection that can, in principle, be arrayed in an infinite list: that is to say, it is possible to describe a systematic listing that sweeps them all out.(The proof is along the lines of the argument that works for the rationals.)

What we have allowed to happen in casually accepting any decimal expansion is to open the door to what are known as the transcendental numbers, those numbers that lie beyond those that arise through euclidean geometry and ordinary

algebraic equations. Cantor's argument shows us that transcendental numbers exist and, in addition, there must be infinitely many of them, for if they formed only afinite collection, they could be placed in front of our list of algebraic numbers(the non-transcendentals), so yielding a listing of all real numbers, which we now know is impossible. What is striking is that we have discovered the existence of these strange numbers without identifying a single one of them! Their existence was revealed simply through comparing certain infinite collections to one another. The transcendentals are the numbers thatfill the huge void between the more familiar algebraic numbers and the collection of all decimal expansions: to borrow an astronomical metaphor, the transcendentals are the dark matter of the number world.

In passing from the rationals to the reals, we are moving from one set to another of higher cardinality as mathematicians put it. Two sets have the same cardinality if their members can be paired off, one against the other. What can be shown using the Cantor argument is that any set has a smaller cardinality than the set formed by taking all of its subsets. This is obvious forfinite collections: indeed, in Chapter 5 it was explained that if we have a set of n members then there are 2n subsets that can be formed in this way. But how large is the set S of all subsets of the infinite set of natural numbers, {1, 2, 3, • • •}?This question is not only interesting in itself but also in the manner in which we arrive at the answer, which is that S is indeed uncountable.

Russell’s Paradox

Suppose to the contrary that S was itself countable, in which case the subsets of the counting numbers could be listed in some order A_1, A_2, • • •. Now an arbitrary number nmay or may not be a member of An- let us consider the set A that consists of all numbers nsuch that n is not a member of the set A_n.NowA is a subset of the counting numbers(possibly the empty subset)and so appears in the aforesaid list at some point, let $A=A_j$ say. An unanswerable question now arises: is j a member of A_j? Ifthe answer were ‘yes’ then, by the very way A is defined, we conclude that j is not a member of A, but $A=A_j$, sothatis self-contradictory. The alternative is no, j is not a member of Aj, in which case, again by the definition of A, weinferthatj is a member of A=Aj, and once more we have contradiction. Since contradiction is unavoidable, our original assumption that the subsets of the counting numbers could be listed in a countable fashion must be false. Indeed, this argument works to show that the set of all subsets of any countable but infinite set is uncountable.

This particular self-referential style argument was introduced by Bertrand Russell(1872- 1970)in a slightly different context that led to what is known as Russell’s Paradox. Russell applied it to the ‘set of all sets that are not members of themselves’, asking the embarrassing question whether or not that set is a member of itself. Again, ‘yes’ implies ‘no’ and ‘no’ implies ‘yes’, forcing Russell to conclude that this set cannot exist.

In the 1890s, Cantor himself discovered an implicit contradiction stemming from the idea of the 'set of all sets'. Indeed, Russell acknowledged that the argument of his paradox was inspired by the work of Cantor. The upshot of all this, however, is that we cannot simply imagine that mathematical sets can be introduced in any manner whatsoever, but some restrictions must be placed on how sets may be specified. Set theorists and logicians have been wrestling with the consequences of this ever since. The most satisfactory resolution of these difficulties is provided by the now standard ZFC Set Theory (the Zermelo-Fraenkel set theory with the Axiom of Choice).

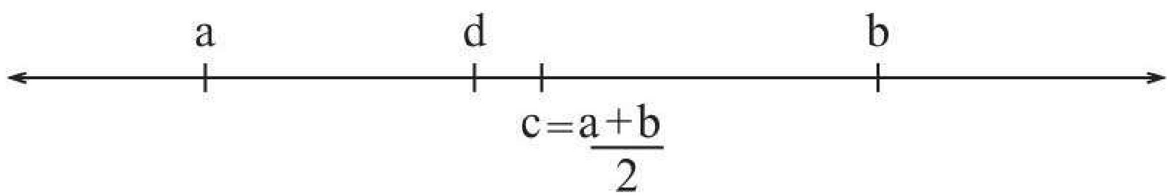
The number line under the microscope

There are different ways of looking at the size of infinite sets of numbers that are revealed if we look at the distribution of the various number types that knit together to bind the number line into a continuum. The rationals may only be a countable collection of numbers, but the collection is densely packed within the line in a way that the integers plainly are not. Given any distinct numbers, a and b, there is a rational number that lies between them. The average of the two numbers, $c = \frac{a+b}{2}$, certainly is a number lying between them, but it may be irrational. However, if c is irrational we can approximate it by a rational number d, witha terminating decimal expansion, by letting d have the same decimal representation as c up√ to a very large number of decimal places. For example, if we take$\sqrt{2}$2=1.414..., wehavethat$\sqrt{2}$

differs from 1.414 by less than 0.001, and each time we take another decimal place we guaranteefinding a rational number that approximates√ 2 more accurately(on average, ten times more accurately)than the previous one. If the number of initial places in which they agree is sufficiently large, then their difference will be so small that both c and d will lie between a and b. The number of places we need to take after the decimal place will depend on just how close a and b are to each other to begin with, but it is always possible tofind a rational d that does the job(see Figure 10). We say that the set of rational numbers is dense in the number line for just this reason. Of course we can, by the same argument, show that there is another rational, splitting the interval from a to d, say, and, inthisway, wearriveatthe conclusion that infinitely many rational numbers lie between any two numbers, however small the difference between these two numbers might happen to be. In particular, there is no such thing as the smallest positive fraction, for, given any positive number, there is always a rational lying between it and zero.

Not to be outdone, the set of irrationals also forms a dense set. Before explaining this, I point out that once we have identified one irrational, the Pythagorean number $\sqrt{2}$ for example, the floodgates open and we can immediately identify infinitely more. When we add a rational to an irrational the result is always an irrational. For example, $\sqrt{2}+7$ is irrational by dint of this fact. In a similar fashion, if we multiply an irrational number by a rational number(other than 0), the result is another irrational number.(Simple

contradiction arguments establish both these claims.)In particular√ , we canfind an irrational number of size as small as we like: $t=\frac{\sqrt{2}}{n}$is irrational for any counting number nand by taking nlarger and larger we can make t as close to 0 as we please.As with the rationals, we therefore see that there is no smallest positive irrational, and hence there is no such thing as the smallest positive number.



10. Rationals separate any given positions on the number line

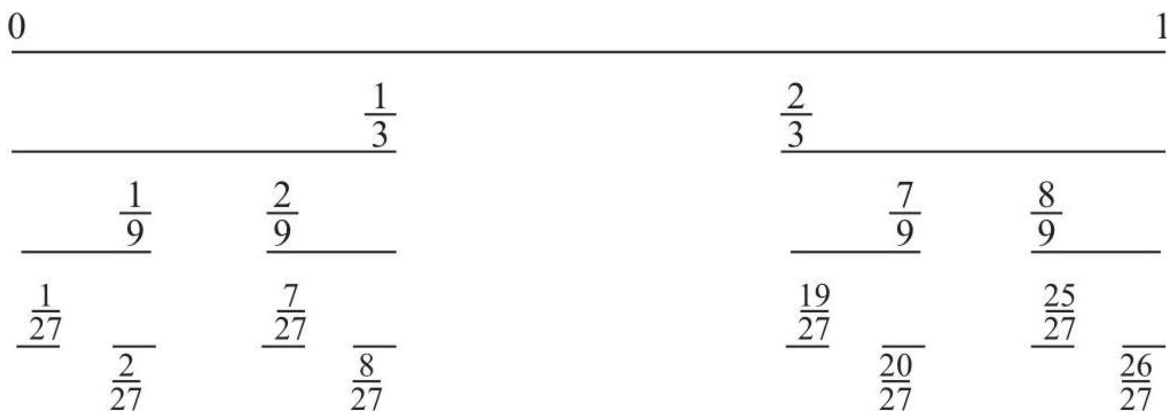
Returning to our given numbers, a and b, once again let c be their average. If c is irrational, we have a number of the required kind(irrational). If on the other hand, c is rational, put d=c+t,where t is the irrational number of the previous paragraph. By what has gone before, d will also be irrational, and if we take nlarge enough,we can always ensure that d is so close to the average c of the two given numbers a and b that it lies between them. In this way, we see that the irrational numbers too form a dense set and, as with the rational numbers, we can infer that there are infinitely many irrational numbers lying between any two numbers on the number line.

And so the set of rationals and its complementary set of irrationals are in one way comparable(they are both dense in

the number line)and in another not(thefirst set is countable, the second not).

Cantor’s Middle Third Set

We now have a clearer idea as to how the rational and the irrational numbers interlace to form the real number line. The rational numbers form a countable set, yet are densely packed into the number line. Cantor’s Middle Third Set is, by way of contrast, an uncountable subset of the unit interval that nevertheless is sparsely spread. It is the result of the following construction.



11. Evolution of Cantor’s Middle Third Set to the 4th level

We begin with the unit interval I, that is all the real numbers from0 up to 1 inclusive. Thefirst step in the formation of Cantor’s set is the removal of the middle third of this interval, that is all the numbers between $\frac{1}{3}$ and $\frac{2}{3}$.The set that remains consists of the two intervals from 0 up to $\frac{1}{3}$and from $\frac{2}{3}$up to 1. At the second stage, we remove the middle

third of these two intervals, at the third stage we remove the middle third of the remaining intervals, and so on. Cantor's Middle Third Set C then consists of all points of I that are not removed at any stage of this process.

The total length of the little intervals that remain as we pass from one stage of this process to the next is, by design, $\frac{2}{3}$ that of the previous stage; it follows that at the nth stage the total length of the surviving intervals is $(\frac{2}{3})^n$. This expression approaches 0 as n increases and since the Cantor set C is the collection of all points that are left at the end of it all, it follows that the 'length', or measure, of C must be 0.

We might suspect that we have thrown the baby out with the bath water and that there are no points at all left in C. Is the Middle Third Set empty?The answer is a resounding no! There are infinitely many numbers left in C. This is best seen if we shift our representation of the numbers of the interval to base three 'decimals' known as ternary, as the whole construction is based on thirds. In base three decimals, the numbers $\frac{1}{3}$ and $\frac{2}{3}$ are respectively given by 0.1and0.2. By discarding the middle third of the unit interval, we have thrown away all those numbers whose ternary expansion begins with 0.1, and indeed the overall process weeds out exactly those numbers that have a 1 anywhere in their ternary expansions. The numbers in C are exactly those whose ternary expansions consist entirely of 0s and 2s.(For example,34survives the infinite cull as in ternary it has the recurring expansion 0.202020....)

Next we make an amazing observation. By taking the ternary expansion of any number c in C and replacing each instance of2 by 1, we obtain the binary expansion of some number c in the unit interval. This gives a one-to-one correspondence of C with the set of all numbers in I(written in binary). It follows that the cardinality of C is the same as that of I, and since the latter is an uncountable set(by Cantor's Diagonal Argument), it follows that the Cantor Middle Third Set is not only infinite but uncountable.

ThereforewehaveasetC that is in one sense negligible in size(has measure zero), but by another way of reckoning C is

huge, as it has the same cardinality as I and hence of the whole real line.

What is more, far from being dense, C is nowhere dense. Recall that by saying that a set like the rationals is dense, we mean that whenever we take a real number a, there are rational numbers to be found in any little interval surrounding a, however small that interval might be. We say that any neighbourhood of a contains members of the set of rationals. The Cantor set has quite the opposite nature- numbers not in C might live their lives in the real line without ever coming across any members of C, provided they confine their experiences to a narrow enough locality around where they live. To see this, take any number a that is not in C, so that a has a ternary expansion that contains at least one 1: a=0•....1....., with the 1 in the nth place, say. For a sufficiently tiny interval surrounding a, the numbers b in that interval have a ternary expansion that agrees with that of a up to places beyond the nth, and so all of them will also not be members of the strange set C as their ternary expansions will also contain at least one instance of 1.

On the other hand, any member a of the Cantor set will not feel too isolated, for when a looks out into any interval J that surrounds it in the number line, however small, a willfind neighbours from the set C living alongside it(and numbers not in C as well). We can specify a member b of the given interval J that also lies in C by taking b to have a ternary expansion that agrees with a to a sufficiently large

number of places, but with no entry being a 1. Indeed, there are uncountably many members of C in J.

In conclusion, the Cantor Middle Third Set C is as numerous as can be and, to the members of the C club, their brothers and sisters are to be seen all around them wherever they look. To the numbers not in C, however, C hardly seems to exist at all. Not one member of C is to be spotted in their exclusive neighbourhoods, and the set C itself has measure zero. To them, C is almost nothing.

Diophantine equations

Some of the principal sets of the number line may be characterized in the language of equations. The rational numbers, which form a countable set, are the numbers that arise as solutions of simple linear equations: the fraction $\frac{b}{a}$ is the solution to the equation $ax-b=0$ (a and b are integers). Numbers like $\sqrt{2}$ that do not arise in this way are called irrational, and they form an uncountable collection that cannot be paired off with the counting numbers in the way that the rationals can. Within the set of irrationals there are the transcendentals, which are the numbers that never arise as the solutions of equations of these kinds even if we allow higher powers of x. It is known that π is an example of a transcendental number, but $\sqrt{2}$ is not, as it solves the equation $x^2-2=0$. The approach then is to define classes of numbers through the kinds of equations they solve.

An interesting line of study emerges, however, when we take the opposite tack and insist that not only the coefficients of our equations but their solutions have also to be integers. Here is a classic example.

A box contains spiders and beetles and 46 legs. How many of each kind of creature are there?This little number puzzle can be solved easily by trial, but it is instructive to note thatfirst, it can be represented by an equation: $6b+8s=46$, and second that we are only interested in certain kinds of solutions to that equation, namely those where the number of beetles(b)and spiders(s)are counting numbers. In general, a system of equations is calledDiophantine when we are restricting our solution search to special number types, typically integer or rational answers are what we are after.

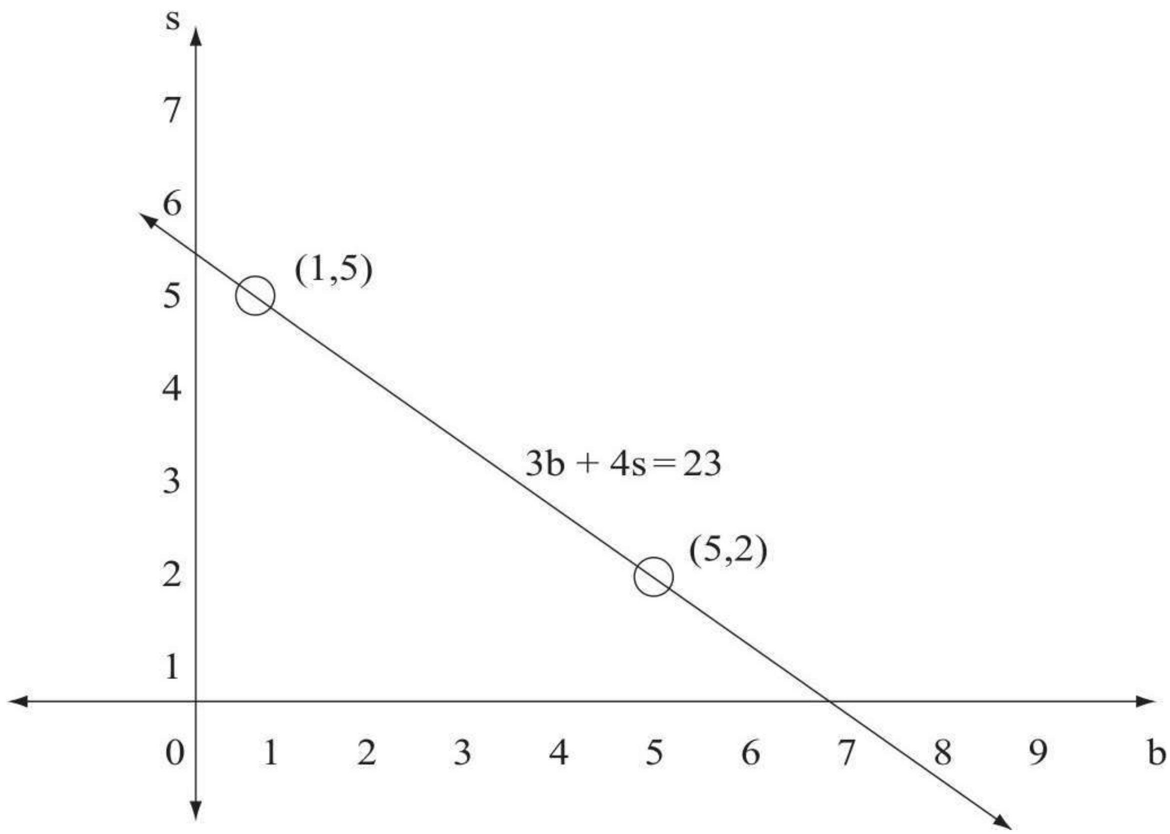
There is a simple method for solving linear Diophantine equations such as this one. First, divide through the equation by the hcf of the coefficients, which are in this case 6 and 8 so their hcf is 2.Cancelling this common factor of 2 we obtain an equivalent equation, that is to say one with the same solutions: $3b+4s=23$. If the right-hand side were no longer an integer after performing this division, that would tell us that there were no integral solutions to the equation and we could stop right there. The next stage is to take one of the coefficients, the smaller one is normally the easiest, and work in multiples of that, in this case 3. Our equation can be written as $3b+3s+s=(3\times 7)+2$; rearranging we obtain $s=(3\times 7)-3b-3s+2$. The point of this is that it shows

that s has the form 3t+2 for some integer t. Substituting s=3t+2 into our equation and making b the subject, we get

$$3b + 4(3t + 2) = 23 \Rightarrow 3b = 15 - 12t \Rightarrow b = 5 - 4t.$$

We now have the complete solution in integers to the Diophantine equation: b=5−4t, s=3t+2. Choosing any integral value for t will give a solution, and all solutions in integers are of this form.

Our original problem, however, was further constrained in that both b and s had to be at least zero, as negative beetles and spiders do not exist. Hence there are only two feasible values of t, those being t=0andt=1, giving us the two possible solutions of 5beetles and 2 spiders, and 1 beetle and 5 spiders. If we interpret the puzzle as meaning that there is a plurality of both types of creature, we have the traditional solution: 5 beetles and 2 spiders.



12. Lattice points on the line of a linear equation

This type of problem is called linear because the graph of the associated equation consists of an infinite line of points. The Diophantine problem then is tofind the lattice points on this line, which are the points where both coordinates are integral or, if we only admit positive solutions, only lattice points in the positive quadrant will do.

However, once we allow squares and higher powers into our equations, the nature of the corresponding problems are much more varied and interesting. A classical problem of this type that has a full solution is that offinding all Pythagorean

triples: positive integers a, b, and c such that $a^2+b^2=c^2$. A Pythagorean triple of course takes its name from the fact that it allows you to draw a right-angled triangle with sides of those integer lengths. The classic example is the (3, 4, 5) triangle. Given any Pythagorean triple, we can generate more of them simply by multiplying all the numbers in the triple by any positive number as the Pythagorean equation will continue to hold. For example, we can double the previous example to get the (6, 8, 10) triple. This, however, gives a similar triangle, one of exactly the same proportions, as the change is only a matter of scale and not of shape. Given the first triangle, we find the second Pythagorean triple simply by measuring the lengths of the sides in units that are half the size of the original units, thereby doubling the numerical size of the dimensions. There are, however, genuinely different triples such as those representing the (5, 12, 13) and the (65, 72, 97) right-angled triangles.

In order to describe all Pythagorean triples, therefore, it is enough to do the job for all triples (a, b, c) where the hcf of the three numbers is 1, as all others are merely scaled-up versions of these. The recipe is as follows. Take any pair of coprime positive integers m and n, with one of them even, and let m denote the larger. Form the triple given by $a=2mn$, $b=m^2-n^2$ and $c=m^2+n^2$. The three numbers a, b, and c then give you a Pythagorean triple (the algebra is easily checked) and the three numbers have no common factor (also not difficult to verify). The three examples above arise

bytakingm=2andn=1inthefirstcase, m=3andn=2inthe second, while for the last triangle we have m=9, n=4. Ittakes more work to verify the converse: any such Pythagorean triple arises in this fashion for suitably chosen values of mand n, and what is more, the representation is unique so that two different pairs(m, n) cannot yield the same triple(a, b, c).

The corresponding equation for cubes and higher powers has no solution at all: for any power $n \geqslant 3$, there are no positive integer triples x, y, andz such that $x^n+y^n=z^n$. This is the famousFermat's Last Theorem, which in future might be known as Wiles's Theorem as it wasfinally proved in the 1990s by Sir Andrew Wiles. Even for the case of cubes, first solved by Euler, this is a very difficult problem. It is, however, relatively easy to show that the sum of two fourth powers is never a square(and so certainly not a fourth power). This is enough to reduce the problem to the case where nis a prime p(meaning that if we solved the problem for all prime exponents, the general result would follow at once), and indeed the problem was solved for so-called regular primes in the19th century. However, the full solution was only realized as a consequence of Wiles settling a deep question called the Shimura- Taniyama Conjecture.

The most intensively investigated Diophantine equation is, however, the Pell equation, $x^2-ny^2=1$, wherenis a positive integer that is not itself a square. Its significance was appreciated very early, for it seems it was studied both in Greece and in India perhaps as far back as 400 BC, because

its solution allows good rational approximations, $\frac{x}{y}$ of $\sqrt{n}$. For example, when n=2, the equation has as one solution pair the numbers x=577and y=408, and $(\frac{x}{y})^2$ =2.000006. The equation is related to the geometric process of what the Greeks called anthyphairesis where one begins with two line segments and continues to subtract the shorter from the longer, a kind of analogue of the Euclidean Algorithm but applied to continuous lengths. Indeed, the Archimedes Cattle Problem mentioned in Chapter 5 leads to an instance of the Pell equation.

Versions of the Pell equation were studied by Diophantus himself around AD 150 but the equation was solved by the great Indian mathematician Brahmagupta(AD 628)and his methods were improved upon by Bhaskara II(AD 1150), who showed how to create new solutions from a seed solution. In Europe, it was Fermat who exhorted mathematicians to turn their attention to Pell' s equation and the complete theory is credited to the reknowned French mathematician Joseph-Louis Lagrange(1736-1813)(the English appellation 'Pell' is an historic accident).

The general method of solution is based on the continued fraction expansion of $\sqrt{n}$.

Fibonacci and continued fractions

Recall the sequence of numbers, 1, 1, 2, 3, 5, 8, 13, 21, • • • discovered by Fibonacci and introduced in Chapter 5. Take a pair of successive terms in this sequence and write the corresponding ratio as 1 plus a

fraction. If we now 'Egyptianize' this fraction by repeatedly dividing top and bottom by the numerator, a striking pattern emerges. Take, for instance

$$\frac{13}{8} = 1 + \frac{5}{8} = 1 + \frac{1}{1 + \frac{3}{5}} = 1 + \frac{1}{1 + \frac{1}{1+\frac{2}{3}}} = 1 + \frac{1}{1 + \frac{1}{1+\frac{1}{1+\frac{1}{1+\frac{1}{1}}}}}$$

We obtain a multi-floored fraction consisting entirely of 1s, and each preceding ratio of Fibonacci numbers appears as we wind through the calculation. This must happen every time: by the very way these numbers are defined, each Fibonacci number is less than twice the next, and so the result of the division will leave a quotient of 1 and the remainder is the preceding Fibonacci number. You will recall that the ratio of successive Fibonacci numbers approaches the Golden Ratio, τ, and so this suggests that τ is the limiting value of the continued fraction consisting entirely of 1s.

As was shown in Chapter 5, the value of an infinite repeating process may be made the subject of an equation based on that process. If we call the value of the infinite fractional tower of 1s by the name a, weseethata satisfies the relation $a=1+\frac{1}{a}$, because what lies underneath the first floor of the fraction is just another copy of a. From this, we see that a satisfies the quadratic equation

$a^2=a+1$, the positive root of which is $\tau = \frac{1+\sqrt{5}}{2}$.

The type of continued fractions that emerge from this process are intrinsically important. When we approximate an irrational number yby rationals we naturally turn to the decimal representation of y. This is excellent for general calculations but,being tied to a particular base, is not mathematically natural.Essential to the nature of y is how well our number y can be approximated by fractions with relatively small denominators. Is there any way tofind a series of fractions that best deals with the conflicting demands of approximating y to a high degree of accuracy while keeping the denominators relatively small?The answer lies in the continued fraction representation of a number that does this through its truncations at ever lower floors.

Continued fractions look very awkward because of the many floors we have used in representing them. However, the inconvenience of writing all thefloors of the division is easily side-stepped- since every numerator is 1, we need only record the quotients in the division to specify which continued fraction we are talking about.For instance, the representation for the fraction $\frac{25}{91}$develops as follows:

$$\frac{25}{91} = 0 + \frac{1}{3 + \frac{16}{25}} = 0 + \frac{1}{3 + \frac{1}{1+\frac{9}{16}}} = \dots$$

that eventually yields a continued fraction specified by the list[0, 3, 1, 1, 1, 3, 2]. As we have seen, the Golden Ratio, τ has the continued fraction

representation[1, 1, 1, 1, • • •]. In a fashion reminiscent of repeating decimal notation, we write $\tau = [\bar{1}]$. The first instance of 3 in the continued fraction of $\frac{25}{91}$ comes from writing 91=3× 25+16, which is thefirst line in the Euclidean Algorithm for the pair(91, 25). Indeed, for this very reason there is one line in the continued fraction for every line of the algorithm when performed on the two numbers. In particular, starting witha reduced fraction in which the two numbers are coprime, the same will apply to each of the fractions that arise in the course of the calculation of the corresponding continued fraction.

The special example afforded by the Golden Ratio opens the door to the idea that we may be able to represent other irrational numbers not byfinite continued fractions(which are obviously just rational themselves)but by infinite ones. But how is the continued fraction of a number a produced? The reader will need to tolerate a little algebraic trickery in order to see this in action, but here is how it goes.

There are two steps in the calculation of a continued fraction for a number $a=[a_0, a_1, a_2, \ldots]$. The number a_0 is the integer part of a, denoted by $a_0=\lfloor a \rfloor$. (Forexample, theintegerpartof $\pi = 3.1415927 \cdots$ is given by $\lfloor \pi \rfloor = 3$.) In general, $a_n=\lfloor r_n \rfloor$, the integer part of rn, where the remainder term rn is defined recursively by $r_0=a$, $r_n=\frac{1}{r_{n-1}-a_{n-1}}$. Applying this for example to $a=\sqrt{2}$ and employing the algebraic device of rationalizing the denominator(with which some readers may be familiar)we obtain, since $\lfloor \sqrt{2} \rfloor = 1$:

$$a = r_0 = \sqrt{2} = 1 + (\sqrt{2} - 1) \text{ so that } a_0 = 1;$$

$$r_1 = \frac{1}{r_0 - a_0} = \frac{1}{\sqrt{2} - 1} = \frac{\sqrt{2} + 1}{(\sqrt{2} - 1)(\sqrt{2} + 1)} = \sqrt{2} + 1,\, a_1 = \lfloor r_1 \rfloor = 2;$$

Next we get

$$r_2 = \frac{1}{r_1 - a_1} = \frac{1}{(\sqrt{2} + 1) - 2} = \frac{1}{\sqrt{2} - 1} = \frac{\sqrt{2} + 1}{(\sqrt{2} - 1)(\sqrt{2} + 1)} = \sqrt{2} + 1,$$

so that $r_1 = r_2 = \cdots = \sqrt{2} + 1,\ a_1 = a_2 = \cdots = 2$, and so $\sqrt{2} = [1, \overline{2}]$.

Indeed, the numbers that have recurring representations as continued fractions are rational numbers(which are exactly the ones whose representations terminate)and those that arise from quadratic equations such as$_\tau$,which we saw above is one solution of the equation $x^2=x+1$,and $\sqrt{2} = [1, \overline{2}]$ which satisfies $x^2=2$.Some other examples showing the rather unpredictable nature of the recurrences are and $\sqrt{28} = [5, \overline{3, 2, 3, 10}]$,There is nevertheless one very particular and remarkable facet to the pattern of the expansion of the continued fraction of an irrational square root. The expansion begins with an integer r, and the recurrent block consists of a palindromic sequence(a sequence of numbers that reads the same in reverse)followed by 2r.This can be seen in all the preceding examples:for instance for $\sqrt{28}$ we see that $r=5$, the palindromic part of the expansion is 3,2,3, which is followed by 2r=10.For $\sqrt{2}$2and $\sqrt{17}$, the palindromic part is empty, but the pattern is still there, albeit in a simple form. It can be shown that the continued fraction

representation of a number is unique- two different continued fractions have different values.

The importance of continued fractions in approximation of irrationals by rationals is due to the so called convergents of the fraction, which are the rational approximations of the original number that result from truncating the representation at some point and working out the corresponding rational number. These represent the best approximation possible to the number in question in the sense that any better approximation will have a larger denominator than that of the convergents. The convergents of the Golden Ratio are the Fibonacci ratios. Since every term in the continued fraction representation of τ is 1, the convergence of these ratios is retarded as much as it possibly could be. For that reason, there is no more difficult number than τ to approximate by rationals and the Fibonacci ratios are the best you can do. If the denominator of a convergent of a continued fraction is q, then the approximation is always within $\frac{1}{\sqrt{5}q^2}$ of the true value of the number and the convergents of a continued fraction alternately underestimate and overestimate the value to which they approach. It is, however, the euclidean numbers such as τ and $\sqrt{2}$ that are the worst when it comes to rational approximation. Some particular transcendentals, whose nature may seem the farthest removed from the rational world, may yet be approximated very closely and have convergents that home in on the target number with great rapidity.

The connection with Pell’s equation, $x^2-ny^2=1$, mentioned at the close of the previous section, now comes about as the solution(x, y)to the equation with the minimum possible positive value of x exists and is to be found among the convergents of the continued fraction representation of $\sqrt{n}$. For example, when n=7the sequence of convergents of $\sqrt{7}$begins with 2/1, 3/1, 5/2, 8/3, • • • and it is x=8, y=3 that provides this smallest so called fundamental solution of the Pell equation $x^2-7y^2=1$. The fundamental solution, however, sometimes does not turn up at all early in the expansion: for example, the smallest positive solution to $x^2-29y^2=1$isx=9801andy=1820. Once this fundamental solution(x, y)has been located, however, all other solutions arise by taking successive powers of the expression $(x+y\sqrt{n})^k$ (k=1, 2, 3, • • •)and extracting the coefficients of the corresponding rational and irrational parts of the expanded expression. In this way, the full solution set of the Pell equa√ tion is realized through the continued fraction representation of $\sqrt{n}$.

Chapter 8 Numbers but not as we know them

Realandcomplexnumbers

The construction of the complex numbers is much simpler and goes much more smoothly than the construction of the real numbers. Thefirst stage in producing the reals is development of the rationals, at which point we have to explain what is meant by a fraction. A fraction, such as 23 is just a pair of integers, which we represent in this familiar but peculiar manner. The idea of fractional parts is not difficult to understand, although the corresponding arithmetic takes real effort to master. Along the way your teachers explain in passing that such fractions as 23,46,69etc. are 'equal' - they are not the same number pairs but they do represent equal slices of pie. This is not hard to accept but it does draw our attention to the fact that a rational number is in reality an infinite set of equivalent fractions, each represented by a pair of integers. This sounds intimidating and we might prefer not to think too much about this, for the prospect of manipulating infinite collections of pairs of integers might leave us feeling uneasy. There is one saving grace in that any fraction has a unique reduced representation where the numerator and denominator are coprime, which can be got by cancelling any common factors in the fraction with which you originally began. Nonetheless, once you are familiar with the properties of fractions and the rules for using them, nothing should go wrong even though

closer examination reveals that, as you do your sums, you are implicitly manipulating infinite collections of integer pairs.

However, things get worse at the next stage when we try to pin down what real numbers truly are. Let' s begin with Pythagoras' problem- he found that there was no fraction equal to $\sqrt{2}$, so we can introduce a new symbol r, endow it with the property that $r^2=2$, and form a new 'field' of numbers from the rationals together with the new number r. This works in that the set of all numbers of the form a+br, wherea and b are rational obey all the normal rules of algebra- we can even divide because the reciprocal of a number of this form retains the form, as can be seen through a little fancy algebraic footwork known as rationalizing the denominator.

The new numbers r and-r furnish the two solutions of the equation $x^2=2$, but what about $x^2=3$? It seems that we need to adjoin yet another new number in order to solve this equation as it is easy to check that no number of the form a+br will square to give 3.(A simple contradiction argument suffices here: assuming that $(a+br)^2=3$, allows you to deduce the false statement that at least one of $\sqrt{2}$ or $\sqrt{3}$ is rational after all.)

It is tempting to cut through all this fretting about particular equations and simply declare that we already know what the real numbers are- they are the collection of all possible decimal expansions, both positive and negative. These are very familiar, in practice we know how to use them,

and so we feel on safe ground. At least until we ask some very basic questions. The main feature of numbers is that you can add, subtract, multiply, and divide. But, for example, how are you supposed to multiply two infinite non-recurring decimals? We depend on decimals beingfinite in length so that you ‘start from the right-hand end’, but there is no such thing with an infinite decimal expansion. It can be done, but it is complicated both in theory and in practice. A number system where you struggle to explain how to add and multiply does not seem satisfactory.

And there are other little pitfalls. When you multiply $\frac{1}{3}$ by 3, the answer is 1. When you multiply 0.333• • •by 3, the answer is surely0.999• • •. It is indeed the case that two different decimal expansions can represent the same number: 1.000• • •=0.999• • •. In fact, this happens with any terminating decimal, for example0.375=0.374999• • •. Hence it can’t be quite right to say that decimals and real numbers are one and the same, as we see that two different decimal expansions can equal the same number. Moreover, the numbers with non-unique decimal expansions will change if we work in another base, and that raises another complication. If we define real numbers by decimals, we are making the construction depend on an arbitrary choice(base ten). If we do the same construction in binary, will the set of ‘real numbers’ be the same? And what do we mean by ‘the same’ in any case?

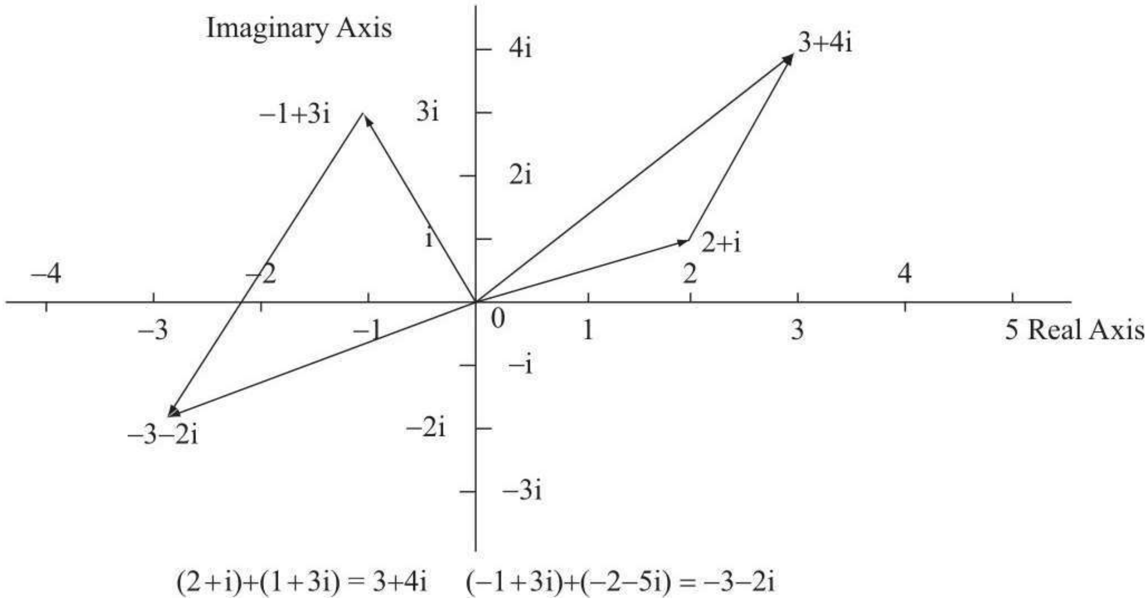
You mayfind the foundational questions raised above interesting or you may grow impatient with all the introspection as we seem to be making trouble for ourselves when previously all was smooth sailing. There is a serious point, however. Mathematicians appreciate that, whenever new mathematical objects are introduced, it important to construct them from known mathematical objects, the way, for instance, fractions can be thought of as pairs of ordinary integers. In this way, we may carefully build up the rules that govern the new extended system and know where we stand. If we neglect foundations completely, it will come back to haunt us later. For example, the rapid development of calculus, which was born out of the study of motion, led to spectacular results, such as prediction of the movement of the planets. However, manipulation of infinite things as if they werefinite sometimes provided amazing insights and at other times patent nonsense. By putting your mathematical systems on afirm foundation, we can learn how to tell the difference. In practice, mathematicians often indulge in 'formal' manipulations in order to see if some sparkling new theorem is in the offing. If the outcome is worthy of attention, the result can be proved rigorously by going back to basics and by invoking results that have been properly established earlier.

This is why Julius Dedekind(1831- 1916)took the trouble of formally constructing the real number system based on his idea that is now referred to as Dedekind cuts of the real line. Thefirst mathematician, however, to successfully deal with

the dilemma caused by the existence of irrational numbers was Eudoxus of Cnidus(fl380 BC)whose Theory of Proportions allowed Archimedes to use the so-called Method of Exhaustion to rigorously derive results on areas and volumes of curved shapes before the advent of calculus some 1,900 years later.

Thefinal piece of the number jigsaw - the imaginary unit

The introduction of a new symbol i that squares to-1is staggeringly successful as it resolves at a stroke not only the problem of providing a solution to one equation but allows solution of all polynomial equations and much more besides. We certainly have two square roots of any negative number,−r,as both the numbers $\pm i\sqrt{r}$ square to−r by virtue of the property that $i^2=-1$ and the ass± umption that, as with ordinary arithmetic,multiplication commutes meaning that zw=wz for any numbers z and w. Indeed, if we continue on the basis that the system of complex numbers a+bi should subsume that of the reals(which correspond to the case where b=0)and that all the normal rules of algebra should continue to be respected, we meet with no difficulties and many pleasant surprises. The set of complex numbers, denoted by C is a 'field' which, among other things,guarantees that division is also possible. To see how it all works out, however, it is best to leave the monorail of the real line and look at life in two dimensions.



13. Addition of complex numbers by adding directed line segments

The arithmetic of complex numbers presents itself very nicely in the complex plane. We think of the complex number a+bi as being represented by the point(a, b)in the coordinate plane.When we add two complex numbers z=(a, b)andw=(c, d), we simply add theirfirst and second entries together, to give us z+w=(a+c, b+d). If we make use of the symbol i,wehavefor example(2+i)+(1+3i)=3+4i.

This corresponds to what is known as vector addition in the plane,where directed line segments(vectors)are added together, top to tail(see Figure 13). We begin at the origin, which has coordinates of(0, 0), and in this example we lay down ourfirst arrow from there to the point(2, 1). To add the number represented by(1, 3),we go to the point(2, 1), and draw an arrow that represents moving 1 unit right in the

horizontal direction(that is the direction of the real axis), and 3 units up in the direction of the vertical(the imaginary axis). We end up at the point with coordinates(3,4). In much the same way, we can define subtraction of complex numbers by subtracting the real and imaginary parts so that, for example,(11+7i)-(2+5i)=9+2i.This can be pictured as starting with the vector(11,7), and subtracting the vector(2,5), tofinish at the point(9,2).

Multiplication is another matter. Formally it is easy to do:we multiply two complex numbers together by multiplying out the brackets, remembering that i2=−1. Assuming the Distributive Law continues to hold, which is the algebraic rule that allows us to expand the brackets in the usual way, then multiplication proceeds as follows:

$$(a+bi)(c+di)=a(c+di)+bi(c+di)=$$

$$ac+adi+bci+bdi^2=(ac-bd)+(ad+bc)i$$

Division, on the other hand, can be calculated by means of thecomplex conjugate. In general, the conjugate of z=a+bi is denoted by $\bar{z}$ and is a−bi, in other words, $\bar{z}$ is the reflection of z in the real axis. The multiplication rule applied to $z\bar{z}$ gives a^2+b^2, which is just a real number as the imaginary part turns out to be zero. This equals the square of the distance of z from the origin, which is denoted by $|z|$. In symbols $z\bar{z}=|z|^2$. We may now divide one complex number by another by multiplying top and bottom by the conjugate of the divisor in order to make the division one by a purely real

number. This is analagous to the standard technique of rationalizing the denominator that is used to remove square roots in the bottomline, which we used to calculate the continued fraction for $\sqrt{2}$. For example:

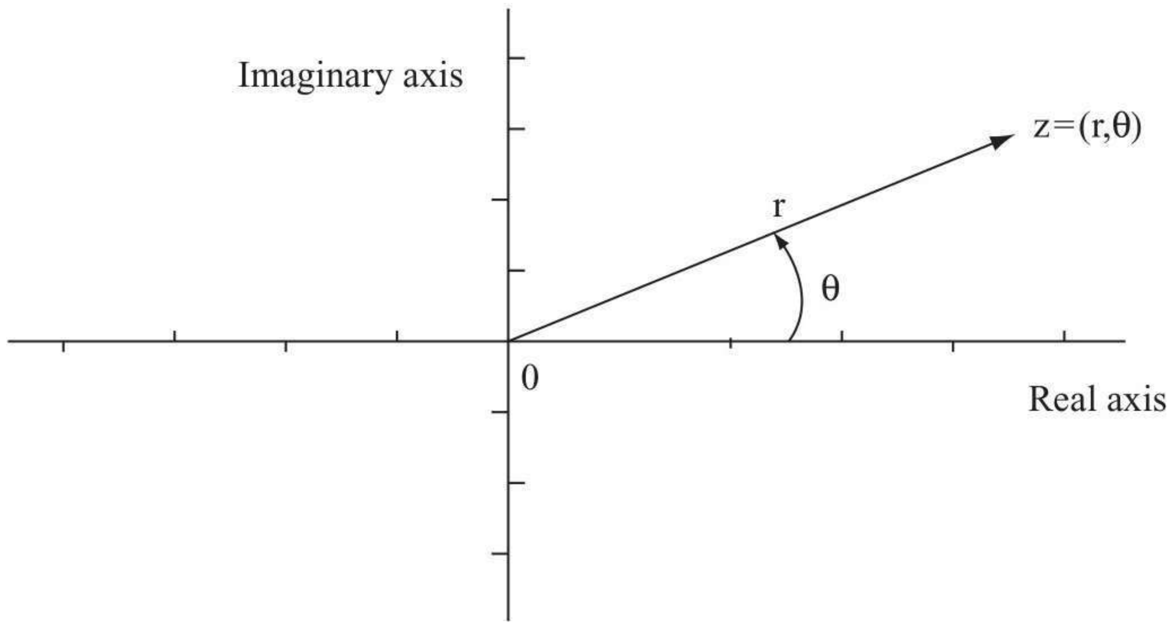
$$\frac{15+16i}{2+3i} = \frac{(15+16i)(2-3i)}{(2+3i)(2-3i)} = \frac{30-45i+32i-48i^2}{2^2-6i+6i-3^2i^2}$$

$$= \frac{(30+48)-(45-32)i}{2^2+3^2} = \frac{78-13i}{13} = 6-i.$$

By using general rather than specific complex numbers we can, in the same way, find the outcome of a general division of complex numbers in terms of their real and imaginary parts as we have done above for general complex multiplication. However, as long as the technique is understood, there is no pressing need to produce and to memorize the resulting formula.

Multiplication has a geometric interpretation that is revealed if we alter our coordinate system from the ordinary rectangular coordinates to polar coordinates. In this system, a point z is once again specified by an ordered pair of numbers, which we shall write as (r, θ). The number r is the distance of our point z from the origin O(called in this context the pole). Therefore r is a non-negative quantity and all points with the same value of r form a circle of radius r centred at the pole. We use the second coordinate θ to specify z on this circle by taking θ to be the angle, measured in an

anti-clockwise direction, from the real axis to the line Oz. The number r is called the modulus(plural moduli)of z,while the angle θ is called the argument of z.



14. The position of a complex number in polarcoordinates

Suppose now that we have two complex numbers, z and w, whose polar coordinates are$(r_1, \theta_1)$and(r_2, θ_2)respectively. It turns out that the polar coordinates of their product zw take on a simple and pleasing form. The rule of combination can be expressed neatly in ordinary language: the modulus of the product zw is the product of the moduli of z and w, while the argument of zw is the sum of the arguments of z and w. In symbols, zw has polar coordinates$(r_1r_2, \theta_1+\theta_2)$. The multiplication of the real numbers is subsumed under this more general way of looking at things: a positive real number r, for instance, has polar coordinates(r,0), and if we

multiply by another(s,0), the result is the expected(rs,0),corresponding to the real number rs.

Much more of the character of the multiplication of complex numbers is revealed through this representation. The polar coordinates of the complex unit i are given by(1,90°). (Normally,angles are not measured in degrees in such circumstances but in the natural mathematical unit of the radian:thereare 2π radians in a circle, so that a turn of one radian corresponds to moving one unit along the circumference of the unit circle, centred at the origin. One radian is about 57.3° .)If we now take any complex number z=(r, θ)and multiply by i=(1,90°), wefind that zi=(r, θ+90°). That is to say, multiplication by i corresponds to rotation through a right angle about the centre of the complex plane. In other words, the right angle, that most fundamental geometric idea, can be represented as a number.

Indeed, the effect of adding or multiplying by a complex number z on all the points in a given region of the complex plane can be pictured geometrically. Imagine any region you fancy in the plane.If we add z to every point inside your region, we simply move each point the same distance and direction determined by the arrow, or vector as we often call it, represented by z.Thatistosaywe translate the region to some other position in the plane so that the shape and size are exactly maintained, as is its attitude, by which we mean the region has not undergone any rotation or reflection.Multiplying every point in your region by z=(r, θ

)hastwoeffects, however, one caused by r and the other by θ. The modulus of each point in the region is increased by a factor r, so all the dimensions of the region are increased by a factor of r also(so its area is multiplied by a factor of r^2). Of course, if r<1 then this 'expansion' is better described as a contraction as the new region will be smaller than the original. The region will, however, maintain its shape- for instance, a triangle is mapped on to a similar triangle with the same angles as before. The effect of θ, as we have explained above, is to rotate the region through an angle θ, anticlockwise about the pole. The net effect then in multiplying all points of your region by z is to expand and rotate your region about the pole. The new region will still have the same shape as before but will be of a different size, determined by r, and will be lying in a different attitude as determined by the rotation angle θ.

Further consequences

The polar version of complex numbers is particularly suited to the taking of powers and roots for to raise z=(r, θ)to some positive power n, we simply raise the modulus to that power, and add θ to itself n times, to give zn=(rn, nθ). The same formula applies to fractional and negative powers. Division can be comprehended in polar form as well. As with real numbers, division by a complex number z means multiplication by its reciprocal w=1z, but what number is this wθGiven that z=(r, θ)the number w is the one with the property that zw=(1,0), the number 1. This shows us that we

must take $w=(r1,-\theta)$, for then $zw=(r,\theta)(r1,-\theta)=(rr1,\theta-\theta)=(1,0)$, as we require. This gives an alternative to the rectangular approach to division that makes use of complex conjugation.

There are a host of applications of complex numbers, even at the elementary level. The interplay between rectangular and polar representations brings trigonometry into play in a surprising and advantageous way. For instance, a standard exercise for students is the derivation of important identities that now arise very naturally by taking arbitrary complex numbers of unit modulus(i.e. $r=1$),and calculating powers using both rectangular and then polar coordinates. Equating the two forms of the answer then reveals a trigonometric equation.

A point with polar coordinates$(1,\theta)$has, by elementary trigonometry, the rectangular coordinates$(\cos\theta, \sin\theta)$. If we now multiply two such complex numbers $z=\cos\theta+i\ \sin\theta$and $w=\cos\phi+i\ \sin\phi$in rectangular coordinates we obtain: equating the real and imaginary parts of the two versions of this one product then painlessly yields the standard angle sum formulas of trigonometry:

$$zw = (\cos\theta\cos\phi - \sin\theta\sin\phi) + i(\cos\theta\sin\phi + \sin\theta\cos\phi)$$

while the same in polar coordinates gives:

$$zw = (1, \theta)(1, \phi) = (1, \theta + \phi) = \cos(\theta + \phi) + i\sin(\theta + \phi);$$

$$\cos(\theta+\phi)=\cos\theta\cos\phi-\sin\theta\sin\phi, \quad \sin(\theta+\phi)=\cos\theta\sin\phi+\sin\theta\cos\phi$$

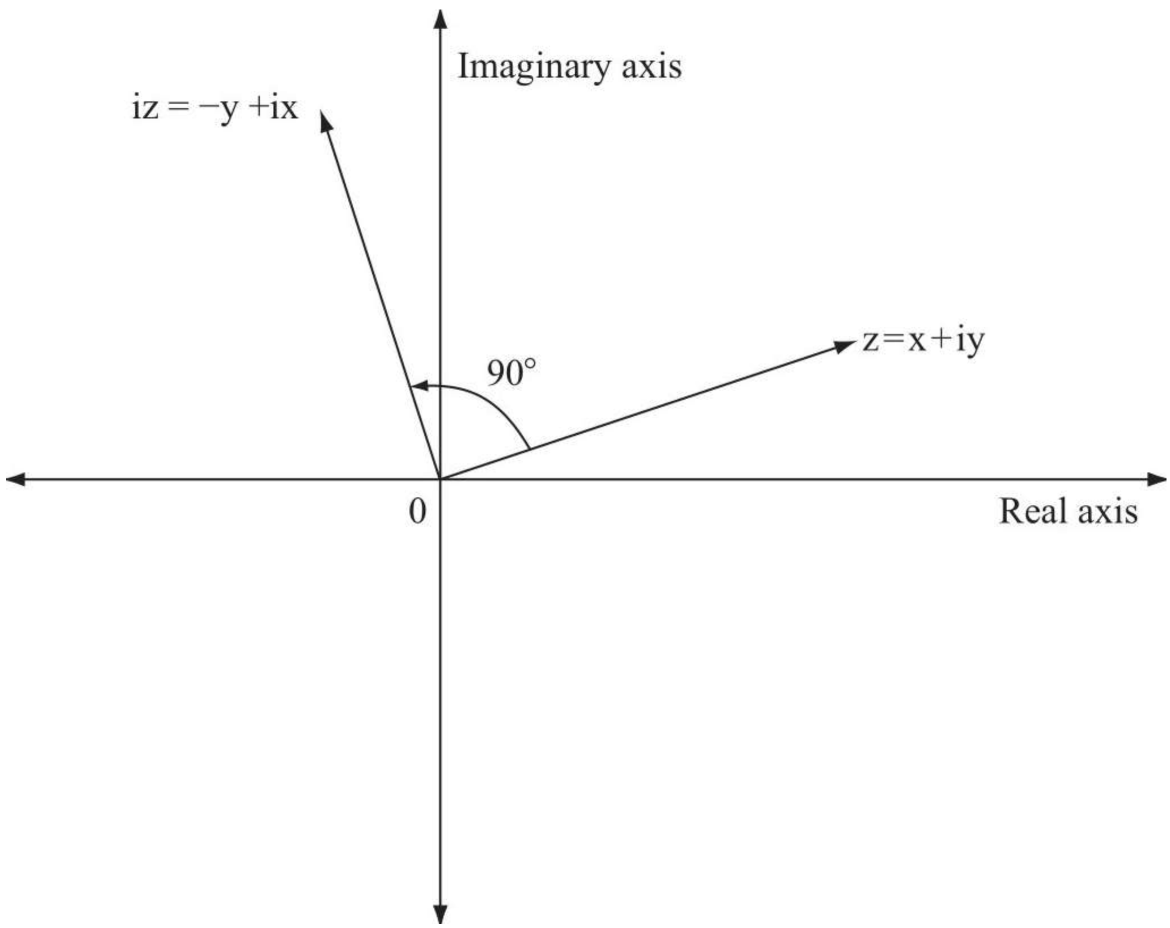
Alternatively, the polar form for complex multiplication can be derived using these trigonometric formulas. Indeed, the rule that we have stated here, without proof, for multiplication in polar form is usuallyfirst derived from the rectangular form by using trigonometric formulas.

Much more comes quite easily now as the use of complex numbers reveals a connection between the exponential or power function, and the seemingly unrelated trigonometric functions. Without passing through the portal offered by the square root of minus one, the connection may be glimpsed, but not understood. The so-called hyperbolic functions arise from taking what are known as the even and odd parts of the exponential function. To every trigonometric identity there corresponds one of identical form, except perhaps for sign, involving these hyperbolic functions. This can be verified easily in any particular case, but then the question remains as to why it should happen at all. Why should the behaviour of one class of functions be so closely mirrored in another class, defined in so different a manner, and of such different character? Resolution of the mystery is by way of the formula $\theta=\cos\theta+i\ \sin\theta$, which shows that the exponential and trigonometric functions are intimately linked, but only

through use of the imaginary unit i.Oncethisisrevealed(foritis surprising and is by no means obvious), it becomes clear that results along the lines described are inevitable through performing calculations using the two alternative representations offered by this equation and then equating real and imaginary parts. Without the formula, however, it all remains a mystery.

Complexnumbersand matrices

Let us examine some consequences of the revelation that multiplication by i represents a rotation through a right angle about the centre of the coordinate plane. If z=x+iy,wehave through expanding the brackets and reordering multiplications that i(x+iy)=-y+ix, so that the point(x, y)istakento(-y, x)under this rotation;see Figure 15. In this way, multiplication by i can be regarded as operating on points in the plane. This operation enjoys the special property that for any two points z and w and any real number a,wehavei(z+w)=iz+iw,andi(aw)=a(iw).Moreover, if we multiply a real number a by a complex number x+iy,wegeta(x+iy)=ax+i(ay). In terms of points in the complex plane, we have that(x, y)ismovedto(ax,ay), or to write it another way, a(x, y)=(ax,ay).



15. Multiplication by i rotates a complex number by a rightangle

The kinds of operations that enjoy these two properties are known as linear and are of paramount importance throughout all mathematics. Here, I wish only to draw to your attention to the fact that the effect of such an operation L is determined by its action on the two points(1, 0)and(0, 1), for let us suppose that L(1, 0)=(a, b)andL(0, 1)=(c, d). Then for any point(x, y)we have(x, y)=x(1, 0)+y(0, 1), and so using the properties of a linear operation we obtain:

$$L(x, y)=L(x(1, 0)+y(0, 1))=xL(1, 0)+yL(0, 1)==x(a, b)+y(c, d)=(ax, bx)+(cy, dy)=(ax+cy, bx+dy).$$

This information may be summarized by what is known as a matrix equation:

$$(x, y)\begin{pmatrix} a & b \\ c & d \end{pmatrix} = (ax + cy, bx + dy).$$

Here we have drawn out an example of matrix multiplication, which indicates how that operation is carried out in general. A matrix is just a rectangular array of rows and columns of numbers. Matrices, however, represent another kind of two-dimensional numerical object and, what is more, they pervade nearly all of higher mathematics, both pure and applied. They represent a whole corpus of algebra, and much of modern mathematics strives to represent itself through matrices, so useful have they proved to be. Two matrices with the same number of rows and the same number of columns as each other are added entry-to-entry: for example, tofind the entry in the second row and third column of the sum of two matrices, we simply add the correspondingly placed entries in the two matrices in question. It is matrix multiplication, however, that gives the subject a new and important character, and how it is conducted has emerged of its own accord in the previous example- each entry in the product matrix is formed by taking the dot product of a row of thefirst matrix with a column of the second, meaning that the entry is the sum of the corresponding products when the row

of thefirst matrix is placed on top of the column of the second.

Matrices follow all the usual laws of algebra except commutativity of multiplication, meaning that for two matrices A and B it is not generally true that AB=BA. However, matrix multiplication is associative, meaning that products of any length may be written unambiguously without the need for bracketing.

Linear transformations of the plane are typically rotations about the origin, reflkections in lines through the origin, enlargments and contractions about the origin, and so called shears(or slanting), which move points parallel to afixed axis by an amount proportional to their distance from that axis in a manner similar to the way the pages of a book can slide past one another. Any sequence of these transformations can be effected by multiplying all of the relevant matrices together to reveal a single matrix that has the same net effect as all those transformations acting in turn. The rows of the resultant matrix are simply the images of the two points(1, 0)and(0, 1), as we saw above, known as basis vectors.

It is now natural to look at the matrix J that represents an anticlockwise rotation of a right angle about the origin as it should mimic the behaviour we see when we multiply by the imaginary unit i. Since the point(1, 0)is taken onto the point(0, 1)by the rotation and similarly the point(1, 0)moves

to(−1, 0), these two vectors form the rows of our matrix J. The result of squaring J will be a matrix that has the geometric effect of rotating points through $2\times 90° = 180°$ about the origin. We calculate this below by matrix multiplication. Tofind, for example, the bottom right entry of J^2 we take the dot product of the second row and second column, which gives $(-1)\times 1+0\times 0=-1+0=-1$. The complete calculation has the following outcome:

$$J^2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I.$$

The matrix I with rows(10)and(01)is the identity matrix, so called as it acts like the number 1 in that when multiplied by another matrix A the result is A. Thematrix−I, which represents a full half turn rotation about the origin, does behave like−1in that $(-I)^2=I$. The upshot of all this is that the matrices aI+bJ, where a and b are real numbers, faithfully mimic the complex numbers a+bi with respect to addition and multiplication, and so give a matrix representation of the complex numberfield. The matrix corresponding to the typical complex number a+bi is

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix}.$$

The matrices that represent the complex numbers do commute with one another but, as was mentioned above, this does not generally apply to all matrix products and another way in which matrices can misbehave is that not all of them can be ‘inverted’.For most square matrices A(a matrix with equal numbers of rows and columns), we mayfind a unique inverse matrix B such that AB=BA=I, the identity matrix. The existence of the inverse matrix however depends upon a single number associated with a square matrix known as its determinant. In general, this is a certain sum of signed products formed by taking one entry from each row and column of the array. For the typical 2×2matrix array as introduced on page 118, the determinant is the number=ad-bc. Determinants have many uses and agreeable properties. For instance, represents the area scale factor of the corresponding matrix transformation:a shape of area a will be transformed into one of area a when undergoing a transformation by that matrix(and if is negative, the shape also undergoes a reflection, reversing the original orientation). What is more, the determinant of the product of two square matrices is the product of the determinants of those matrices. A square matrix A will have an inverse B except in the case where=0,inwhich case it will not. A zero determinant corresponds geometrically to a degenerate transformation where areas are collapsed by the matrix tofigures of zero area such as a line segment or even a single point.

For the matrix of a complex number $z=a+bi$, wenotethat=a^2+b^2, whichisneverzeroexceptwhenz=0- but of course the number 0 never had a reciprocal before, and that remains the case in the wider arena of the complex numbers. This does confirm however that every non-zero complex number possesses a multiplicative inverse.

We stand here on the edge of the vast worlds of linear algebra, representation theory, and applications to multi-dimensional calculus, and this is not the place to go further. However, the reader should be aware that matrices apply to three dimensions and indeed to n-dimensional space, typically through $n\times n$ matrices. Although the arrays become larger and more complicated, the matrices themselves yet remain two-dimensional numerical objects.

Numbersbeyondthecomplexplane

Thefield C of all complex numbers is complete in two important ways. An infinite sequence of complex numbers in which the terms cluster into ever smaller circles of radius that approaches 0 is called convergent. Any convergent sequence of complex numbers approaches a limiting complex number. This is also true of the real numbers, but not of the rationals- the successive decimal approximations to any irrational number represent a sequence of rational numbers that approach a limit outside of the rationals. Moreover, C is complete (or closed) in the algebraic sense that it can be shown that any polynomial equation

$p(z)=a+bz+cz^2+\cdots+z^n=0$ has n (complex) solutions, z_1, z_2, $\cdots$, z_n, which then allows $p(z)$ itself to be fully factorized as $p(z)=(z-z_1)(z-z_2)\cdots(z-z_n)$.

This and other stunning successes of the complex numbers largely obviate the need to expand the number system further beyond the complex plane. Indeed, it is not possible to construct an augmented number system that contains C and also retains all the normal laws of algebra. Moreover, there are only two extended systems that retain much algebraic structure at all, these being the quaternions and the octonions. Although their use is not nearly so widespread as that of the complex numbers, the quaternions are put to work, for example, in three-dimensional computer graphics. The octonions, which can be thought of as pairs of quaternions, lack not only the commutative property but also the associative property of multiplication.

A quaternion is a number of the form $z=a+bi+cj+dk$, where thefirst part $a+bi$ is an ordinary complex number and the two quaternion units j and k also satisfy $j^2=k^2=-1$. In order to do multiplication with quaternions, we need to know how the units multiply with one another and this is determined by the rules $ij=k$, $jk=i$, $ki=j$ but the reversed products carry the opposite sign, so that, for example, $ji=-k$ (indeed, all these products may be derived from the single additional equation: $ijk=-1$). The quaternions then form an enhanced algebraic system that satisfies all the laws of algebra except for commutativity of multiplication, due to the sign

changes mentioned above in the reversed products. The consistency of the system can also be demonstrated through representation by 2× 2 matrices, but this time we allow complex rather than just real entries. The number 1 is once more identified with I, the identity matrix but the units i, j, andk have as their matrix counterparts:

$$i = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \; j = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \; k = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

while the typical quaternion z hasasitsmatrix:

$$z = \begin{pmatrix} a + bi & c + di \\ -c + di & a - bi \end{pmatrix}.$$

This representation of the quaternions by matrices is not unique, however, and indeed the representation of the complex numbers by matrices also has equivalent alternatives. Moreover, it is possible to represent the quaternions without employing complex numbers but only at the expense of using larger matrix arrays: the quaternions can be represented by certain 4× 4matriceswith only real number entries.

New kinds of numbers and the extensions of old systems have come about through the need to perform calculations the outcome of which could not be accommodated by the number

system as it stood. Every civilization begins with the counting numbers, but calculations involving fragments lead to fractions, those involving debt lead to negatives, and as Pythagoras discovered, those involving lengths lead to irrational numbers. Although a very ancient revelation, the fact that not all numerical matters could be dealt with using whole numbers and their ratios was a subtle discovery of a deeper kind. As science became more sophisticated, the number systems required have needed to mature in order to deal with these advances. Scientists do not generally look to create new numbers systems in a whimsical fashion. On the contrary, they are introduced often reluctantly and hesitatingly atfirst, to deal with research problems. For example, althoughfirst introduced in the 19th century, matrices arose irresistibly in quantum mechanics in the early 20th century when physicists encountered a quantity of the form $q=AB-BA$that was nevertheless not zero. In any commutative system of numbers, q would of course be 0, so the numerical objects needed here were not of a kind they had met before: they were matrices.

It seems now that the world of mathematics and physics has enough number types. Although there are kinds of numbers not mentioned in this book, the number types that are commonly used throughout mathematics and science have not needed to change a great deal since thefirst half of the 20th century.

These observations, however, bring our mathematical balloon ride to its conclusion. We began at ground level and have ascended to where I hope the reader can gaze down upon a view of the rich and mysterious world of numbers.

Further reading

Two other books in the OUP VSI series that complement and expand on the current one are Mathematics by the Field’s medallist Timothy Gowers and Cryptography by Fred Piper and Sean Murphy. Probability and statistics, fields that were neglected here in Numbers, are the subject of the VSI Statistics by David J. Hand.

An insight into the nature of numbers can be read in David Flannery’s book, The Square Root of 2: A Dialogue Concerning a Number and a Sequence(Copernicus Books, 2006). This leisurely account is in the Socratic mode of a conversation between a teacher and pupil. One to Nine: The Inner Life of Numbers by Andrew Hodges(Short Books, 2007)analyses the significance of thefirst nine digits in order. Actually it uses each number as an umbrella for examining certain fundamental aspects of the world and introduces the reader to all manner of deep ideas. This contrasts with Tony Crilly’s 50 Mathematical Ideas You Really Need To Know(Quercus Publishing, 2007), which does as it says, digesting each of 50 notions into a four-page description in as straightforward a manner as possible. The explanations are mainly through example with a modest amount of algebraic manipulations involved, rounded off with historical details and timelines surrounding the commentary. A particularly nice account on matters concerned with binomial coefficients is the paperback of Martin Griffiths,

The Backbone of Pascal's Triangle(UK Mathematics Trust, 2007), in which you will read proofs of Bertrand's Postulate and Chebyshev's Theorem, giving bounds for the number of primes less than n.

Elementary Number Theory by G. and J. Jones(Springer-Verlag, 1998) gives a gentle but rigorous introduction and goes as far as aspects of the famous Riemann Zeta Function and Fermat's Last Theorem. The classic book An Introduction to the Theory of Numbers, byG.H.HardyandE.M.Wright, 6thedn(Oxford University Press, 2008) assumes little particular mathematical knowledge but hits the ground running. The author's book Number Story: From Counting to Cryptography(Copernicus Books, 2008) has more in the way of the history of numbers than this VSI and includes mathematical details in thefinal chapter. The Book of Numbers by John Conway and Richard Guy(Springer-Verlag, 1996) is full of history, vivid pictures, and all manner of facts about numbers. Quite a lot of the history and mystery surrounding complexnumbers is to be found in An Imaginary Tale: The Story of $\sqrt{-1}$ (Princeton University Press, 1998) by Paul J. Nahin. Paul Halmos' s Naive Set Theory(Springer-Verlag, 1974) gives a quick mathematical introduction to infinite cardinal and ordinal numbers, which were not introduced here.

A popular account of the Riemann Zeta Function is the book by Marcus du Sautoy, The Music of the Primes, Why an Unsolved Problem in Mathematics Matters(HarperCollins, 2004),

while Carl Sabbagh's, Dr Riemann's Zeros(Atlantic Books, 2003)treats essentially the same topic.

There are two accounts of the solution to Fermat's Last Theorem, those being Fermat's Last Theorem: Unlocking the Secret of an Ancient Mathematical Problem by Amir D. Aczel(Penguin, 1996)and Fermat's Last Theorem by Simon Singh(Fourth Estate, 1999). The best popular book on the history of coding up to the RSA cipher is also an effort of Simon Singh: The Code Book(Fourth Estate, 2000). The unsolvability of the quintic(fifth-degree polynomial equations)was not explained in our text here but is the subject of an historical account: Abel'sProof: AnEssayonthe Sources and Meaning of Mathematical Unsolvability(MIT Press, 2003)by Peter Pesic.

Websites

A very high-quality web page that allows you to dip into any mathematical topic, and is especially rich in number matters, is Eric Wolfram's MathWorld: mathworld.wolfram.com. For mathematical history topics, try The MacTutor History of Mathematics archive at St Andrews University, Scotland: www-history.mcs.st-andrews.ac.uk/history.index.html. Webpages accessed 8 October 2010. Wikipedia's treatment of mathematics by topic is generally serious and of good quality, although the degree of difficulty of the treatments is a little variable. For

example, Wikipedia gives a good quick overview of important topics such as matrices and linear algebra.

牛津通识课：概率

[英]约翰·黑格 著
徐大成 译

东方出版中心

图书在版编目（CIP）数据

牛津通识课. 概率 / (英) 约翰 · 黑格著 ; 徐大成译. -- 上海 : 东方出版中心 , 2020.12
ISBN 978-7-5473-1744-0

Ⅰ. ①牛… Ⅱ. ①约… ②徐… Ⅲ. ①科学知识－普及读物②概率－普及读物 Ⅳ. ①Z228②O211.1-49

中国版本图书馆CIP数据核字（2020）第248911号

上海市版权局著作权合同登记：图字09-2020-1022号

Probability: A Very Short Introduction was originally published in English in 2012. This translation is published by arrangement with Oxford University Press. Dook Media Group Ltd is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

中文版权：©2021 读客文化股份有限公司

牛津通识课：概率

著　　者　[英] 约翰 · 黑格
译　　者　徐大成
责任编辑　江彦懿
封面设计　汪芝灵

出版发行　东方出版中心
地　　址　上海市仙霞路345号
邮政编码　200336
电　　话　021-62417400
印 刷 者　天津联城印刷有限公司

开　　本　787mm x 1092mm　1/32
印　　张　7
字　　数　80千字
版　　次　2021年2月第1版
印　　次　2021年2月第1次印刷
定　　价　32.00元

版权所有　侵权必究
如图书有印装质量问题，请寄回本社出版部调换或电话010-87681002联系。

目录

致谢

01 基本原理 Fundamentals

02 概率的运作 The Workings of Probability

03 历史概要 Historical Sketch

04 概率试验 Chance Experiments

05 理解概率 Maki ng Sense of Probabi l ities

06 人们玩的游戏 Games People Play

07 在科学、医学和运筹学中的应用 Appl ications in Science, Medici ne, and Operations Research

08 其他应用 Other Appl ications

09 有趣而棘手的问题 Curiosities and Di lemmas

附录 书中提出的问题的答案

术语对照

Contents

Acknowledgements

List of illustrations

Chapter 1 Fundamentals

Chapter 2 The workings of probability

Chapter 3 Historical sketch

Chapter 4 Chance experiments

Chapter 5 Making sense of probabilities

Chapter 6 Games people play

Chapter 7 Applications in science, medicine, and operations research

Chapter 8 Other applications

Chapter 9 Curiosities and dilemmas
Appendix
References and further reading

致谢

我要感谢莱顿·沃恩·威廉姆斯(Leighton Vaughan Williams)和迈克·史密斯(Mike Smith)，因为他们提供给我关于赛马赔率的信息；我要感谢莱斯·米勒(Les Miller)提供他对轮盘赌的仿真结果；我要感谢伊恩·麦克海尔(Ian McHale)对如何估计足球世界杯决赛圈中32支队伍的胜率的详尽解释；我还要感谢德里克·罗宾逊(Derek Robinson)在图表方面对我的帮助。我要特别感谢一位匿名评论者，他鞭辟入里、富有建设性的建议帮助我消除歧义，更连贯地编排内容和提取这个学科的中心思想。

对那些认为自己是书中的信息、观点和逸事的来源，但未被我提及的人，我恳求原谅，我引用了太多资料以至于不能列出全部出处。这不是一本其中每个断言都能找到其来源的学术专著，本书意在提升读者们对概率这门学科是什么、如何发展和如何应用的兴趣。

所有书中遗留的谬误和缺陷都来自我自己。

约翰·黑格

2011年8月

01 基本原理 Fundamentals

概率的视角

概率是不确定性这一概念的形式化表述。误打误撞效应显然到处都是。从生物学上说，我们都是父母基因随机混合后的产物。像是石油泄漏、火山喷发、海啸、地震等灾害，或是中彩票这样令人愉悦的事情，都会随机且显著地影响人们的生活。

许多人具有良好的理解概率的直觉，但在你对某件事情有了某种先入为主的观点，而后来一些具有不完全明显的相关性的新事实被披露出来的时候，这种理解就会让你误入歧途。的确有一些臭名昭著的有关生日、二孩家庭、有三个选择的电视节目游戏的“诡计问题”(trick questions)，它们似乎被设计成说服你这门学科是有违常识的。其实概率并不违背常识，只要清除掉或者考虑到这些问题中所有隐藏的假设，合理的答案就会浮出水面。只不过概率的确需要清楚的思维过程。

概率的广泛应用促进了这门学科概念和方法的发展。1944年6月的诺曼底登陆[1]能够发生，就是因为当时人们认为有利天气出现的概率相对较高。荷兰的工程师们在建造保护其国家免受海洋侵袭的堤岸时，必须考虑发生严重洪水的概率。一种新型治疗方法是否比先前的方法更能帮助一名患者多生存五年？你需要交多少钱来给自己、车辆、房子或财产上保险取决于早期索赔的可能性。你所做的大多数决定：在学校学习什么、选择谁作为人生伴侣、在哪里居住、从事什么工作都是在有不确定性的情况下进行的。就像皮埃尔-西蒙·拉普拉斯(Pierre-Simon Laplace)在1814年所说的那样：

> ……生命中最重要的问题大多都只是概率问题。

“概率是……”这样的措辞无论何时出现，都伴随着某些假设(它们可能在不经意间被忽略了)。如果那些假设是无端的，那么这些断言就不会被人相信。我希望在这本书中假设是明确的，无论它们是含蓄还是直白。在我们将目光转向概率的种种阐述能如何被诠释之前，先描述一下产生这些阐述的不同思路。**古典概率**

概率的**古典**(classical)或者说**客观**(objective)视角经常出现在有关概率的游戏中，例如掷色子和转轮盘赌。这些过程都会产生一系列可能的结果，我们出于对称性的考虑，或者因为找不到是其中一个结果而不是另一个更会发生的原因，认为它们都是等可能的。所以我们只是对结果进行计数，并赋予它们相等的概率，这样试验中的任何事件的概率都被认为是引发它的结果占所有结果的比率。

例如，连掷两次硬币，四种可能的正反面结果是：正正、正反、反正、反反。就一枚**公正的**硬币来说，每次掷出正或反都是等可能的，所以四个结果中没有一个比另一个更可能或更不可能，每一个结果的概率都应该是1/4。其中有三个至少一次掷出正面，所以总体上讲正面出现的概率是3/4。

从一个牌堆中取两张扑克牌，有1326种结果(请相信我的话)。如果牌堆是被洗好了的，我们就认为这些扑克牌组合都是等可能的。因为其中有64种牌面由一张A和一张“十牌”(即10、J、Q或K[2])组成，所以我们得出结论，抽到这样的组合——“二十一点(Blackjack)”——的概率是64/1326，刚好不到5%。

仅从概率的角度而言，这些例子都可以转化为从装有完全相同的球的袋子中取出某个球的形式。第一个例子对应的袋子中装有4个球，3个是红球；第二个例子对应的袋子中装有1326个球，其中64个是红球。的确，每一个对概率的客观考量的例子本质上都与从袋子或者瓮

中取出一个球的问题完全相同(这就解释了学生们教材中这类例子过多的原因)。

我要强调的是，仅仅计算可能结果的数量然后计算多少个结果会引发相应的事件是不够的。一定要有令人信服的理由说明任何结果都不会比其他的更可能或更不可能发生。否则，基于彩票只有两个可能的结果：要么中奖，要么不中，你会掉入买彩票中大奖概率是50%的思维陷阱中!

试验证据——频率

我们希望在“大富翁”这类家庭游戏或者例如双色子赌博的赌场游戏中，色子的六个面中掷出每一个都是等可能的。但如果色子由不均匀的材料制成，或者它的长度、宽度和高度三者不相同，那么假定每种结果是等可能的显然不明智。在相同条件下进行的一系列投掷过程中，出现任何一个面的频率都会波动，但最终将会稳定并趋近于一个特定值。

不可能出现前1000次投掷中20%的结果是6点，而接下来的1000次投掷中这个比例跳到了60%。在这些可重复试验中，结果可能是不完全一样的，但是每一个结果都倾向于表现出某个特定的频率，**频率论者**(frequentist)认为这个频率值就是相应结果的概率。

对于一个不完美的色子，在前1000次投掷中，我们可能会得到170次6点，下1000次中，可能得到181次6点，诸如此类。我们不能从这些试验中推断出掷出6点的概率**精确值**，但是试验数据指导我们对概率进行估计，我们收集的试验数据越多，我们估计得就越准确。我们无法知道概率的精确值，但这一事实并不能否认概率的存在。

如果我从洗好的牌堆中抽取一张牌，似乎没有理由认为某种花色比其他花色更容易被抽到。每种花色都有1/4的客观概率。而且如果我放回这张牌，重新洗牌，然后再进行100次试验，我会预期每种花色的出现是同样的频繁，就是大约25次。类似地，对于投掷结果都是等可能的普通色子，投掷结果是5点的概率客观地讲是1/6。在600次投掷中，我们预期掷出5点的次数大约为100次。

在重复大量具有等可能性结果的试验时，任何特定结果相应的**频率**都预期会接近于它客观计算的概率。一个公正的硬币 极少会在100次投掷中给出50次正面朝上的结果，但是直觉上我们不知道该期望投掷结果多么接近理想情况才合理。

频率观点不仅被应用于同样条件下的重复性试验，还有在即将出生的婴儿是男是女上。不考虑家庭因素，我们来看看从许多国家和文化环境中收集的覆盖了很长时间跨度的数据。一个持续的模式是：每49个女婴出生，就有51个男婴出生。鉴于无法将某个新生儿和其余的进行区分，一个频率论者会认为生男孩的概率是51%。

一些规模惊人的试验已经开展了。1894年，动物学家拉斐尔·韦尔登(Raphael Weldon)发表了将12个色子投掷2600次的结果。他的数据与六个面等可能出现的观点相抵触，因为5和6这两个数字出现得太频繁。为了辨认数字，他的色子上每个面钻了小孔，刻有5和6的面分别对着刻有1和2的面。这些色子的重心就会更接近数字较小的面，这给出了一个对观察结果中频率过大貌似正确的解释。

大约70年后，一个有大量时间的一丝不苟的人——威拉德·朗克尔(Willard Longcor)在哈佛大学顶尖的统计学家弗雷德里克·莫斯特勒(Frederick Mosteller)手下效力。在莫斯特勒的指导下，朗克尔收集了超过200个色子，并将它们中的每一个都投掷了超过20 000次，只记录结果的奇偶性——得到超过400万个数据。为了让每次投掷的环境

尽可能相同，他使用了一个铺了毯子的桌面，用一个升起来的台阶将色子弹下去。那些类似韦尔登使用的廉价色子存在微小但明显的偏差，以至于出现了太多的偶数，这并不出人意料，也是那些钻孔的原因。而对于那些使用在拉斯维加斯赌场的高精度色子，上面表示数字的点不是轻轻画上去的就是极薄的圆盘贴上去的，就没有可检测到的偏差。这些色子各种结果的频率与在古典视角下等可能结果的概率是一致的。

“二十一点”专家皮特·格里芬(Peter Griffin)挖苦地说，他在拉斯维加斯玩的1820局牌中，庄家牌堆顶上要么是十牌，要么是A的情况出现了770次。而抽到这些对庄家有利的牌的客观概率是5/13，所以格里芬怀疑自己是否被欺骗了——随机概率只会让发牌者抽到这种好牌大约700次。

2002年3月，马拉维有6202名五岁以下的儿童被认为疑似患上了肺炎，其中523名儿童死亡，死亡率为8.4%。已知没有某些特殊情况让这段时期不同于以往，一个频率论者就会推断：一名患上肺炎的马拉维儿童的死亡率是8%~9%。从客观角度来说，关于马拉维患有肺炎的儿童的死亡率的一般性陈述仍是一种推测，尽管基于这样确凿的证据：如果随机从那些**特定的**6202名儿童中选择一名，他的死亡概率是8.4%。

我们将会在后面更深入地讨论频率数据和客观概率的关系。

主观诠释

布鲁诺·德·菲内蒂(Bruno de Finetti)是概率这个领域中最有影响力的思想者之一，他曾写过：

> 概率不存在。

作为概率理论方面的教授，他并不是在将自己研究的学科比作海市蜃楼，而是在驳斥例如“正面朝上的概率是1/2”这种**绝对性**的陈述。对于他来说，每一个包含概率的陈述都是观点的表达，这种表达基于一个人自己的经验和知识，并且有可能在更多的信息被发现的时候发生变化。

考虑如下五个断言：

英国板球队队长会在下一次国际板球对抗赛猜对硬币；

奥斯卡金像奖最佳男主角奖获得者，无论是谁，都会在下一年再次获奖；

没有奥斯陆出生的人曾经获得过奥运会击剑金牌；

理查三世(Richard III)应该对“塔中王子[3]”的死负责；

如果拉尔夫·纳德(Ralph Nader)没有成为候选人，阿尔·戈尔(Al Gore)本会在2000年被选为美国总统。

对于这其中的每一个推断，我们都能够给出自己的**可信度**(degree of believe)、**个人概率**

(personal probability)或者**主观概率**(subjective probability)。这将会是一些非负数，并且不大于1，就是说它是一个介于0%和100%(含)之间的百分比。

0和1分别代表着两个极端——**不可能**和**必然**。我确信在本世纪内足球世界杯必然会再次由非洲国家举办。我认为年龄小于20岁的人不可能获得诺贝尔物理学奖[4]。

评估主观概率

上面的五个断言各具有不同的性质，关于它们我们有多种不同的佐证。对第一个断言来说，我们能用正面和反面的对称性加以反驳；对第二个断言来说，我们可以参考1929年以来的奥斯卡奖历史记录，

前两个情形都能在很短时间内确定其真实与否；第三个断言，无论是真是假，都可以通过盘点奥运会获奖记录来确定；第四个或真或假，但我们永远都无法确定；我们不能让历史重来去探明第五个断言是真是假。

后面会有一些具体的例子来阐释主观概率是如何被评估的。除了这些观点之外，有至少三个一般性评估主观概率的不同方法。一个就是做出一个事件发生与否的合理赌注。但是这个方法不总是对每个人有用：有些人原则上抵制赌博，还有一些人不考虑进行可能导致个人损失的行为。而且对于那些愿意赌博的人来说，他们的**合理赌注**也可能会随着他们站在打赌双方的哪一边而变化。

第二个评估某件事可信度的方法就显得有些刻意了。你会选以下哪一个：猜某一个事件是否发生，或者猜牌堆顶上第一张牌的颜色是红色还是黑色，猜对了获得5英镑。如果你更喜欢后者，就说明你认为此事件的可信度在50%以下。

假设我们继续比较如下两种情况的预期，这个事件是否发生，还是猜第一张牌的花色，猜对获得5英镑。后者有25%的可能会发生，所以你对这两种情况的选择，会反映你认为这个事件的可信度是比25%低，还是在25%到50%之间。

更加精细地比较这些数值会让我们无法确定更偏好哪一边。你对这件事的可信度就会接近于那个选牌的客观概率。你也许会想要使用装有20或者100个完全一样的球的罐子来明确地评估这个事件的可信度，而不是计算分数很困难的有52张牌的牌堆。

这里给出一些具有合适精确度的结果。2010年，网球运动员约翰·伊斯内尔(John Isner)和尼古拉斯·马胡(Nicolas Mahut)进行了温布尔登网球锦标赛[5]史上最长的比赛。经过计算，他们在下一年再次

成为对手(这的确发生了)的精确概率是2/285，或许应该四舍五入到“略低于1%”。但是《 星际迷航》 [6]的一集中，史波克先生(Mr. Spock)告诉柯克(Kirk)他们逃脱的胜算“大约是7824.7比1”，这就很荒谬了。

第三种方法，想一笔金额大小合适的钱，别太少以至于对你来说无关紧要(比如一便士)，也别太多以至于拥有了它就会对你的境遇产生巨大的影响(对大多数人来说是100万，对于比尔·盖茨就要数额大一些了)。我觉得10英镑就行——把它叫作单位金额。

现在假设，不知何故某个事件的真实与否会在明天揭晓：并且如果这个事件是真实的你会获得这个单位金额，如果它是假的就什么也得不到。但有一种提议是：不用等到明天，今天你会获得单位金额中确定比例p的一部分，但对你来说今天或者明天得到这笔钱没有什么差别。

如果p特别小，你可能就会拒绝这个提议，并且更愿意等待；如果p接近于单位1，你可能会接受提议中这个确定比例的金额。但是中间会存在一个p的值让你在接受这个提议与等待明天结果被揭晓中摇摆不定。这个p就是你认为这个陈述或者事件的可信度。

这里我提供我自己的对上述五种断言的主观答案。我认为没有合理的原因来解释为什么一方比另一方更可能在板球掷硬币 中获胜，所以给出的第一个数字是50%；浏览奥斯卡奖的历史，不仅是演员奖，其他类别的奖项也只是零星地在相邻年份中重复颁发——可能现在参选人更多了，所以我给出3%，或者更低；挪威人不以善于击剑著称，但是重剑、花剑、佩剑这些击剑项目自1896年以来一直出现在所有的夏季奥运会中，一些奥斯陆出生的人也许曾经在某次获得过金牌，但是我很怀疑——这里我给出的数字大约是95%；出于对白玫瑰郡[7]的偏

爱，而不是客观证据，对于第四个断言我给出10%；对于第五个断言，考虑到每一个州的投票情况和纳德获得的貌似合理的票数，我给出20%。

在这儿停一下，给出你们自己对这五个断言的意见。在事情不确定的时候，你越善于评估概率，你在生活中做的决定就越可能让你开心。

赔率

无论我们使用古典概型、频率诠释还是可信度，**赔率**(odds)这个词语在描述概率的时候经常出现。我们可能会说用公正的色子掷出6的赔率是“1赔5”——在一系列投掷中，每一次我们掷出6，预期都会有五次掷不出。如果一个结果预期比它的反面更有可能发生，例如排名更高的网球选手获得比赛的胜利，那这个结果就被叫作**有胜算的**(odds on)。

概率和赔率之间有确定的对应关系，我们能够简单地将其相互转换。思考一下频率将会很有帮助。如果概率是20%，或者说1/5，我们预期这个事件在五次机会中发生一次，所以赔率是“1赔4”。如果概率是75%，我们预期它会在四次中发生三次，给出“3赔1”的赔率。如果赔率是“5赔6”，这就表明每五次事件发生，就会有六次事件不发生，所以概率是5/11。

你不必拘泥于数字。在洗好的牌堆顶上的牌是K或者Q的概率是2/13。这可以被表述为“2赔11”，或者同样精确的“1赔5.5”。喜欢哪个就用哪个。

虽然短语“赔率是1赔1”从来不被使用，但是它很有含义。它表明期望一个事件发生和不发生的机会是五五开，所以它的概率是1/2。

然而，我们会板着脸说“赔率是均衡的”。

需要解决的问题

对于如何使用概率，我们没有重大的分歧，但是我们曾经讨论过的三种方法的信徒们可能会从不同的角度分别论述它们的价值。每一个观点都有其用途。为了理解这门学科的运作方式，无论从哪个思路我们都要探寻适当的观点。

客观方法被限制于有限多个结果的情况，所有的这些结果都被判断为等可能的。但是没有硬币或者色子是完美地对称的，基于什么我们可以把这些不完美当作无关紧要的元素而不去考虑它们呢？甚至于我们能否确定可能结果的数量呢？例如假设我们有一个装有两个球的罐子，这两个球要么均为白色，要么均为黑色，要么一白一黑。我们是否可以说有3种等可能的情况，或者球在按顺序被置入的时候，实际上是白白、白黑、黑白或者黑黑这4种等可能情况？持不同看法的人会对两个球均为黑球的概率给出不同的答案。或者假设你到达了一条分出三条岔路的路口，两条路通往新城，第三条路通往海港，做一个“随机选择”：你去往海港的概率是1/3(三个出口中的一个)还是1/2(两个目的地中的一个)？

一个频率论者希望处理可重复试验，它在完全相同的情况下能够不限次数地进行。试验结果的数量不需要是有限的——想想掷同一枚硬币直到正面连续出现3次，或者在一根棍子上取一个随机的点。但是，无论我们多么小心，试验环境都不会保持**绝对**一致，而且任何极限值都只能做估计。怎样描述这种估计中的误差？宣称误差在2%以下的概率是99%，就会引入循环论证——我们需要知道概率是多少，以便定义它!

一个国家入侵另一个国家的概率，或者特定的一次心脏移植成功的概率，这类问题中的情形只出现一次，而且备选结果不能被减少为有限列表中的等可能的结果。客观和频率方法对这些事件都无能为力。这就需要主观概率了。

主观主义者必须确保她相信的事情都是自洽的。例如，在英国国家彩票[8]中，一个机器从列表{1， 2， 3，…， 49}中选取6个数字，苏西也许会倾向于认为约1400万种选择均是等可能的。那么，当问到下面哪一个更有可能的时候：

(a)抽取的数字中没有超过44的；

(b)那些抽取到的数字中不包括连续数字。

她或许会在想了一会儿之后选择二者中的某一个。但是只要她选择了这些事件中的**任何一个**，她都会因为自己的观点不能自洽而愧疚，因为合理的计算显示这两个事件发生的可能性正好是相等的! 对于这种不自洽性，主观概率方法仅仅要求它被解决，但是并未给出确切的解决方式。

因为相比具有有限多种等可能选择的情形，我们希望考虑更宽泛的情况，在考虑不能不限次数地重复试验的情况时，我们将主观概率方法作为默认选项。而且一旦有客观或者频率方法的支持，我们将会更加坚信我们的观点。

解读

借用“袋子中的球”的视角，一些事件的概率被当作是袋子中红色球的比例。所以仅当袋子中没有红色球的时候，概率的值才是0。在这种情况下，这个事件永远不会发生。类似地，概率为单位1对应着每一个球都是红色，所以这种情况下这个事件每次都会发生。只有0和单

位1这些值，才可能确凿地被试验证据证明是**错误的**：如果事件发生了，它的概率就不可能是0；当它没有发生，它的概率就不可能是单位1。而且这从频率或者主观方法来说也是对的。假设概率是某些中间值，比如说3/4。

我们首先来处理一个十分细致的问题。无论一个轮盘赌轮被设计得多么好，从本质上讲所有标着数字的格子被转到的概率**精确**相同是不可能的。赌场需要的是这些概率足够接近理想情况，以至于不大可能分辨出任何数字的概率比其他数字更多或者更少。类似的说法也适用于色子、硬币 和纸牌。所以类似于“概率是3/4”的说法，意味着对于所有实际目的来说概率都足够接近3/4。否则，一个书呆子就会沾沾自喜地告诉你，他**知道**概率不是3/4，并且不害怕引起争执。

在可重复试验的背景下，我们期望从这个断言中得到什么信息呢，“得到红球的概率是3/4”？值得强调的是，我们并不会期望如果进行4次这个试验(每次取球之后放回)，我们会精确地在其中3次抽到红球。可能的情况是，4次重复试验根本没抽到红球，或者甚至每次都是红球。但是在一系列漫长的重复之后，我们的确期望红球出现的频率接近3/4。

漫长的重复试验有多长，或者结果需要多么接近3/4？没有一个确定的、非黑即白的答案。如果在前40次重复试验中，我们只有20次抽到红球，我会强烈质疑概率是3/4的断言，但是如果接下来的40次中得到28次红球的结果，那些质疑就会被削弱。相信或者不相信这个断言可能会在相当长的一段时间内是临时立场。假设试验条件的确一直保持不变，我们使用**所有**收集到的数据来做决定——试验次数过少会引起误导。

稍后我会提供一些准则，并且证明它们。我们以重复100次试验为例，假设概率是一个中间值，接近一半。计算这个数字与由数据得到

的真实频率的**差值**：如果这个差值超过0.1，我会对这个断言产生一些怀疑；如果差值超过0.15，我会产生强烈质疑。在重复试验1000次而不是100次后，我期望结果有更强的一致性，所以用0.03和0.05代替原来的数字。如果假想的数字接近0或者单位1，比如说10%或者90%，我也会期望更强的一致性。在重复试验的基础上，特定的概率更容易让人信服，而**不是**某个所谓的值。

对于一个主观评估，例如明天降雨的概率是60%，情况是怎样的呢？我们不能数百次地再现今天的天气情况，然后检查降雨是有多频繁。这种“试验”只能够进行一次。但是我们也许可以通过检查这个数字产生的过程来检验这个断言。天气预报员使用天气规律的模型来得到他们的结论，即使他们的电脑屏幕上的数字是31.067%，他们也会聪明地给出大约的数字，你会听到“降雨的概率是30%”。所以现在你就能收集不同日期的数据，看看经验证据——在去年给出降水概率为30%的83天中，有多少天真的下雨了？只要那个比例合理地接近于30%，你对这个方法的信心就会增强，所以接受对“明天”的降雨概率就是个理性回应了。

概率是在不确定情况下做决定的关键。如果你真诚地相信特定的一件事或者一个论断的概率是单位1，那么你应该按照它无疑会发生一样来行事；如果你真诚地相信概率是0，那么就按照它好像绝不会发生一样来行事。

如果你认为概率是从0到1之间的某个值，那么就按照你预期它会发生的比例来行事。例如，如果你的判断是概率为60%，想象你会面临这样的情况100次，在60次中(但是你不知道是哪60次)这个事件会发生，而在40次中不会发生。努力理解，考虑到这种权衡，决定你的行动。如果你猜测概率是80%，说明你预期这个事件会更频繁地发生，你的行动可能就会不同。

就像大主教约瑟夫·巴特勒(Bishop Joseph Butler)1736年在他的《 宗教的类比》(Analogy of Religion)中写的那样：“对于我们来说，概率正是生活的准则。”

[1] 诺曼底登陆(The D-Day Invasion of Normandy)是第二次世界大战时西方盟军在欧洲西线战场发起的一场大规模攻势，为“霸王行动”的一部分。

[2] 这四种牌面都算作10点，原文为“ten-card”。

[3] 塔中王子(the Princes in the Tower)，指英格兰国王爱德华四世(Edward IV)的两个儿子：爱德华五世(Edward V)和约克公爵(Duke of York)。他们被理查三世关进伦敦塔之后失踪。

[4] 原文为：“I think it is impossible for someone under twenty years of age to win a Nobel Prize.”本书成书于2014年马拉拉·优素福·扎伊获得诺贝尔和平奖之前，经与原书作者John Haigh沟通，此处添加“诺贝尔物理学奖”的限定。

[5] 温布尔登网球锦标赛(The Championships, Wimbledon)，网球运动中历史最长和最具声望的公开赛之一。

[6] 《 星际迷航》(Star Trek)，美国系列科幻娱乐影视剧。

[7] 指约克郡，白玫瑰为其与约克王朝的共同象征。理查三世为约克王朝最后一任国王。

[8] 英国国家特许经营的彩票，开始于1994年。

02 概率的运作 The Workings of Probability

除了主观的、客观的和频率的理解方法，还有其他理解概率的视角。例如，一定要坚持将一个概率对应于某一个数字吗？我们是否可以说一个概率更大，或者一件事情的可信度比另外一件事情更高？我们真的必须提出一组公理——不言而喻的事实——并据此建立一套理论？

许多杰出的作者都认为建立两个独立的理解概率的方法是有用的，一个是可信度，另一个是古典概率。两者应该具有相同的逻辑规律，不自相矛盾，但两者对于概率是如何生成的和被理解的可以不同。任何理论都应该与古典观点一致，基于可重复实验都会给出等可能的结果，所以我们将着眼于这些案例，寻找概率必须遵循的规则。

加法定理

从洗好的牌堆里面取一张牌。我们认为抽到所有牌都是等可能的，所以求出任何事件的概率——例如抽到梅花、黑桃或者A——就是计算这些事件占总事件的比例。我们如何求出两个事件之中的**每一个**发生的概率呢？

如果两个事件的所有可能结果中没有任何相同，我们称这两个事件是**相互排斥**(mutually exclusive)或者**不相容**(disjoint)的。“抽到黑桃”和“抽到梅花”这两个事件是不相容的，但是“抽到黑桃”和“抽到A”这两个事件不是，因为“抽到黑桃A”同时属于这两个事件。当两个事件**互斥**，这两个事件中任何一个发生的结果总数就是其分别发生的结果数之和，所以我们有一个简单的结论：

当两个事件互斥时，至少一个事件发生的概率是两个事件各自发生概率的和。

这就是概率的**加法定理**(the Additional Law)。它显然适用于所有我们能够以古典视角观察的试验：用袋子中的球作类比，这个定理可以被理解为抽中红或蓝球的结果总数是红球的总数和蓝球的总数之和。而且在任何可重复试验中——例如掷色子或者旋转轮盘赌轮——两个不相交事件的频率和一定是至少一个事件发生的频率。所以加法定理从频率角度看也成立。

一个持主观视角的人也能接受这个定理。否则，存在两个互斥事件，称为A和B，加法定理对它们不成立。这种情况下，主观主义者会面临三个赌局：一个赌事件A发生，一个赌事件B发生，一个赌事件A和B至少一个发生。而对他来说每个赌局都是公平的，他会接受，但他如果参与全部三个赌局，则必定会输钱。加法定理就禁止了这样的矛盾。

加法定理可以拓展到包含大量事件的集合中，前提是这些事件中任意两个都没有相同的结果——它们是**两两不相交的**(pairwise disjoint)。即使一个事件包含了1 000 000种不同的结果，其发生的概率也仅仅是每种结果单独的概率的和。但是假设结果的个数不再有限，例如连续掷一枚硬币 直到正面出现时掷硬币 的次数。

这个试验可能的结果组成一个无限长的表{1， 2， 3， 4，…}，表中的每一个数值都对应着其非零的概率。正面出现时掷硬币 次数为**偶数**的概率是多少？在{2， 4， 6， 8，…}中的结果会让这个事件发生。我们能通过计算其概率和来计算这个事件的概率吗？

数学上这个加法计算没有很大困难，但这个操作已经超出了古典概型的范围，古典概型只能处理**有限**多个可能的结果。这种**无限**长的表中的事件概率的加法定理是不是概率在起作用，人们对此还没有达成共识。有利于将其包括到概率的范畴的是：我们也许可以得到更多种类事件的概率。不利于将其包括到概率的范畴的是：对这个事件的

概率计算不是古典理论的一部分，我们需要在计算中小心会带来陷阱的步骤。这个问题没有正确或者错误的答案。

我是一个实用主义者，我满意于加法定理这样的拓展应用，而且我从没有对这样拓展带来的结果失望过。这种态度是大学中讲授这门学科的大多数教材中给出的标准诠释。但是德·菲内蒂从谨慎的角度建议避免进行这种拓展，有一部分人也这样认为。

乘法定理

掷一枚普通的硬币，你会预期猜对正面或者反面的次数是总次数的一半。洗牌之后预测牌堆顶的牌是红色或者黑色，你也会估计猜对的次数是总次数的一半。当你同时猜测掷硬币 的结果和牌堆顶牌的颜色，有多大可能两个**都**正确？

假设做这种双重试验100次。预期你大约50次猜对掷硬币 的结果，当你猜对之后，预期你继续在一半的次数中猜对牌的颜色。这意味着你大约有25次两个都猜对，看起来得到了25%或者1/4作为两个都猜对的概率。对这样的试验来说，两个都正确的概率就是将两个事件单独成立的概率相乘。

10个大小和材质均相同的球被标有数字0到9，它们中的任何一个被抽中都是完全随机的。所以球上写着较小数字(0~4)或者较大数字(5~9)是等可能的。其中5个数字用绿色写成，另外5个用蓝色写成，所以绿色和蓝色也是等可能的。我们猜颜色或猜数字较小还是较大都分别有50%的概率。那么球上既标有较小数字又是绿色数字的概率是多少？

前文关于硬币 和扑克牌的论断意味着答案是1/4，但是想一会儿你就能发现这不对。在有10个球的情况下，不可能其中的1/4(2.5个)是

较小的绿色数字! 正确的答案取决于哪一些数字是绿色的，哪一些数字是蓝色的。不妨假设1至5是绿色，其余是蓝色。

这种情况下，10个数字里面有4个(1、2、3和4)是较小的绿色数字，所以前述事件的概率是0.4。但就像我们处理第一个问题时一样，我们也可以使用两步走的过程：100次重复试验中，我们预期得到较小数字50次。5个里面有4个较小数字是绿色的，所以在我们得到较小数字的情况下，我们预测绿色次数占4/5。总体上讲，我们预期得到较小的绿色数字40次，再次指向了0.4这一答案。

在硬币和扑克牌的问题中，掷硬币的结果对扑克牌的结果没有影响。我们不会因为得知了硬币掷出正面就改变头脑中抽到红色牌的概率——在**给定**第一个事件发生的条件下，另一个事件的**条件概率**(conditional probability)就是它的正常的概率。如果这成立，这两个事件被称为是**独立的**(independent)，两个事件同时发生的概率就是两个事件单独发生的概率的乘积。

对于10个球的问题，两个事件同时发生的概率也作为乘积出现，其中第一个乘数是一个事件(较小数字)的概率，而第二个乘数是得到较小数字时得到绿色数字的条件概率。所以两个计算在形式上是相同的，唯一的不同就是第一个事件的结果会否影响第二个事件。这两个计算中，我们都用到了概率的**乘法定理**(the Multiplication Law)：

> 两个事件同时发生的概率就是第一个事件发生的概率与第一个事件发生时第二个事件发生的概率的乘积。

独立性

我们用“独立的”这个术语来描述一种情况：第一个事件的发生并不影响我们对第二个事件概率的评估。假设这是成立的。但假如我

们知道了**第二个**事件已经发生，这可能影响我们对第一个事件概率的评估吗？

不会。一个事件的发生与否不会影响另一个事件发生的概率，第二个事件是否发生也并不会影响第一个事件的概率。当两个事件中任何一个发生与否均不会对另一个事件的概率产生影响时，这两个事件是独立的。要计算两个事件同时发生的概率，就将它们各自的概率相乘。

彼此不相互影响的事无疑是独立事件，例如突尼斯今天下雨和巴黎新生儿的性别。但有时独立性并不明显。使用一个公正的普通色子，考虑事件“得到偶数”和“得到3的倍数”，它们的概率分别是1/2和1/3。只有得到6的时候两个事件同时发生，概率是1/6。因为1/2和1/3相乘等于1/6，这两个事件是独立的。得到偶数的概率并不会在我们得知是否得到3的倍数之后改变(反之亦然)。

现在，当你有一个8面色子或者10面色子的时候，考虑相同的问题，色子的每个面都分别被标记了1~8或者1~10。再进行相应的算术过程：你会发现在其中一种情况下两个事件**是**独立的，但是在另一种情况下**不是**。判断独立性时，直觉是有用的，但是并不总是足够的。

在两个因素并不独立的时候假设它们是独立的，是评估概率过程中最常犯的错误。假设在一所大学的研究生院中一半的学生是女生，并且1/5的学生学习工程学科。随机选择一个学生：这个学生是女生的概率会被认为是1/2，这个学生学习工程的概率会被认为是1/5。然而你会发现这个学生是个女工程师的概率远远小于这两个值的乘积——1/10。

具有重叠的事件

加法定理说明了如何计算两个事件中至少一个发生的概率，只要这些事件是互斥的。如果二者不互斥会怎么样？例如，随机抽取一张卡片，是黑桃或者A的概率是多少？黑桃A同时属于这两个分类，所以如果我们只是将各自的概率相加，我们就会将黑桃A计算两次。为了计算两个事件中**至少一个**发生的概率，并纠正可能会被重复计算的结果，就将各自的概率相加，然后减去两个同时发生的概率。

如果两个事件是互斥的，就不可能同时发生，所以这个额外项的值为0，我们就回到了原先的加法定理。

让我们来看看在先前两个例子中这种观点的实际应用。在猜硬币和红牌黑牌的问题中，我们至少猜对一个的概率来自计算过程1/2 + 1/2 – 1/4，等于3/4。在另一个例子中，随机地抽取有标号的球，是较小数字或者绿色的概率是1/2 + 1/2 – 0.4 = 0.6。

抽到黑桃或者A的概率就是13/52 + 4/52 – 1/52 = 16/52，这可以被证明，因为52张牌中刚好有16张满足条件。

最后一个计算过程会警示你不要提前进行算术简化。的确13/52与1/4相等，4/52与1/13相等，但是要将1/4和1/13相加，你最好用原始的分数。将一个像5/13这样的美观的分数写成它的丑陋的近似小数(0.384 615 38……)几乎没有好处。

两个以上的事件

由许多事件组成的集合中的一些事件发生与否，并不影响任何其他事件的概率时，这些事件就被描述为独立的。在这种情况下，乘法定理意味着，以任何方式从这个集合中选取事件，它们均发生的概率仅仅是它们各自的概率的乘积。

但是在3个或者更多的事件**不是**独立的时候，我们如何得到它们均发生的概率呢？例如惠斯特纸牌和桥牌[1]游戏中，纸牌都被随机洗好然后等量地分给四个玩家。他们全都恰好抽到一张A的可能性有多大？

考虑4个单独的事件：甲恰好抽到一张A，乙恰好抽到一张A，丙恰好抽到一张A，丁恰好抽到一张A[2]。显而易见，这4个事件并不是独立的，因为如果任意3个均发生的时候，最后一个就一定发生了。我们会通过一个**三段式**过程来得到它们均发生的概率。

首先，我们计算甲恰好抽到一张A的概率。假设所有可能的分牌方式都是等可能的，我们有了一个计算练习：计算可能的分牌方式的总数，然后计算甲恰好抽到一张A的方式有多少种。相信我，这个概率算下来略低于44%。

假设甲只有一张A(因此有12张非A)。这就给其他的玩家留下了3张A和36张非A，同时乙随机地抽取了其中的13张。对这个较小的牌堆，相似的计数过程指出乙恰好抽到一张A的概率略低于46%。乘法定理告诉我们，两个事件，即甲和乙均恰好抽到一张A发生的概率是这两个值的乘积，略高于20%。

现在我们假设甲和乙每个人都恰好有一张A。然后丙从剩下的两张A和24张非A中随机地取13张牌：他恰好得到一张A的概率是52%。

最后一步就是再一次使用乘法定理，将最后的两个计算过程组合起来：甲、乙和丙均恰好抽到一张A的概率略高于10%。如果这发生了，丁就不可避免地得到最后一张A，所以我们找到了我们所寻求的答案。

虽然分牌是随机的，最公平的A的分配却是颇不可能的。这个答案本身没有什么实际意义，但我们使用的**方法**是具有普遍意义的。为了得到集合中每一个事件均发生的概率，就应该将整个过程分成若干阶段。找到一个事件的概率；然后，**假设**这个事件真的发生，再找到第二个事件的概率；再**假设前两者**均发生，找到第三个事件的概率；然后**假设前三者**均发生，找到第四个事件的概率——以此类推。最终，将所有的数据相乘。

在其他情形中我们是否也要依照这个过程？假设我的交通行程有三个阶段，并且我可以评估它们各自不延迟的概率：但所有阶段都会受天气影响，并且一个阶段延迟与否会改变其他阶段延迟的概率。在制造业中，一件生产设备的安全依赖于数个并不独立工作的组件——其中的一些可能使用相同的供水系统，另一些可能由同一个不可靠的员工做了不充分的测试。在手术中，可能会出现问题的事情和其他事情是否独立会对全部过程能够顺利进行的概率产生巨大的影响。

如果事件是相互独立的，那么它们全都发生的概率就只是它们各自概率的乘积。但我们很少会足够幸运地处于这种情况，实际上在一个分段的评估中，随着工作的推进概率在不断变化，这是一种常态。

三个或者更多的事件中**至少一个**发生的概率是多少呢？加法定理的确可以拓展到这种情况，然而因为这个表达式实在是难以处理，我不会在这儿写下来。它的推导过程与之前描述过的将乘法定理应用于许多事件均发生时所使用的过程一样：一步一步来。

将**独立**当作**互斥**是常见的错误，反之亦然。随机地抽取一张牌的例子会帮助你认识到如何避免它。这里，“抽到一张黑桃”和“抽到一张梅花”这两个事件是互斥的，但绝不是相互独立的，因为如果二者之中任何一个发生了，另一个就不可能发生，所以两个同时发生的

概率是0。同时“抽到一张黑桃”和“抽到一张A”是相互独立的(是吗？)，但显然不是互斥的。

记住：加法定理用来计算至少一个事件发生的概率，乘法定理用来推导它们全部发生的概率。

有时人们说：计数真的只有1、2、无穷大。这个说法揭示了一条真理，如果我们可以完成从处理一件事到处理两件事的过渡，那随后到第3、4、5等的过渡相比而言就不那么重要了。这个道理当然是对加法定理和乘法定理都成立的。

一个巧妙的把戏

任何事件要么会发生，要么不会发生。总概率被分成了事件发生和不发生两部分。所以如果我们能够找到事件不发生的概率，把这个概率从100%中减掉，就能够推断出它发生的概率。

举例来说，我们要计算掷两次公正的色子时，至少得到一个6的概率。任何结果都写成表示第1次和第2次的投掷结果的一对数字，例如(5， 2)或(4， 4)，并且我们认为所有这样的结果都是等可能的。每次投掷都会有6个可能的结果，以至于总共有6× 6=36种结果。我们的事件在没有色子是6的时候不发生，一共5× 5=25种情况。没有6的概率是25/36，所以至少一个6的概率是11/36，比1/3略小一点。

这就引出了1654年布莱瑟·帕斯卡(Blaise Pascal)和皮埃尔·德·费马(Pierre de Fermat)解决的一个点数问题的初级版本。我们必须掷多少次色子才能使我们至少掷出一次6是**更有可能**的，即掷出一个6的概率比1/2大？我们刚刚看到，掷两次是不够的。

每一次额外的投掷都会让可能的结果数增加6倍，同时未掷出6的结果数乘以5。所以3次投掷一共有216种结果，而且其中125种(超过一

半)没有包含6，3次投掷也是不够的。然而4次投掷得到1296种结果，并且其中只有625种是未掷出6的，少于一半。这就使得包括6的结果多于不包括6的结果，所以这时包含6更有可能。4次投掷就足够了。

实际上帕斯卡和费马分析的游戏中不只包含了掷一个色子，还包含了同时掷两个色子的情况；并且设问若要使**两个6**同时出现至少一次更有可能，需要多少次重复同时投掷两个色子。解法是相同的，但是原始的计算过程是艰难的。如今我们可以借助小型计算机或者袖珍计算器来快速地得到结果，然而直到17世纪，较便捷的对数和计算尺才被使用。直到第24次投掷没有双6产生的可能性都更大，但是第25次投掷就将扭转这一局面。

大多数具有“计算这些事件中至少一个发生的概率”这种格式的问题，都可以用这种方式解决：计算它们均不发生的概率，然后从单位1中减掉这个概率。

[1] 惠斯特纸牌(whist)和桥牌(bridge)均为经典的纸牌游戏。

[2] 原文中，甲乙丙丁分别为：Anne、Brain、Colin和Debby。

03 历史概要 Historical Sketch

开端

在1600年左右的佛罗伦萨，有一种关注三个普通色子总点数的游戏。所有色子掷出1(即总点数为3)和所有色子掷出6(即总点数为18)这两种情况出现得最少，其他大多总点数都接近于这个范围的中间值。你应该能发现得到9点有6种方法(例如6 + 2 + 1、5 + 2 + 2，等等)，得到10点也是有6种方法。通常认为，这就“应该”使色子总点数为9和10出现的频率一样。但是一段时间之后，玩家们注意到总点数为10出现得比9明显多。他们就此向伽利略(Galileo)请教一个解释。

伽利略指出他们计数的方法有缺陷。将色子涂成红色、绿色和蓝色，并按涂色的顺序列举出结果。从3 + 3 + 3得到总点数为9需要三个色子具有相同的点数，只有一种方式能够使其发生，(3， 3， 3)。但是5 + 2 + 2的组合可以通过(5， 2， 2)、(2， 5， 2)或(2， 2， 5)中产生，所以这个组合倾向于出现得比前者频繁3倍；6 + 2 + 1通过(6， 2， 1)、(6， 1， 2)、(2， 6， 1)、(2， 1， 6)、(1， 6， 2)和(1， 2， 6)产生，所以这个组合有6种途径产生。一个合理的寻求不同总点数出现频繁程度的方法需要考虑这种因素，而且这种因素确实使得获得10点比9点有更多的方式。佛罗伦萨的赌徒们(Florentine gamblers)学习了关于概率的重要一课——一定要学会**正确地**计数。

1654年夏天，帕斯卡(在巴黎)和费马(在图卢兹)就**点数分配问题**(the problem of points)进行了一次通信。假设史密斯和琼斯约定进行一系列的比赛，首先赢得3局的是获胜者；但不幸的是，当史密斯领先琼斯的比分为2∶1时比赛必须中止。该如何分配赌金？

那时这样的问题已经被提出了至少150年了，仍没有令人满意的解答，但帕斯卡和费马各自独立地找到了一个解决方案，对任意的目标得分和任意的比赛意外终止时的比分，都能够在两人之间**公平**地瓜分赌金。他们使用了不同的方法，但是得到了相同的结果，两人都对对方的才华表示赞赏。对上述具体的问题，应该按照3：1的比例分配，史密斯得到3/4的赌金，琼斯得到1/4的赌金。

他们解法的关键是假设在未来的任何对局中两个玩家获胜是等可能的。他们就每一个玩家计算了能够使其获得最终胜利的可能的假想对局结果数量，并提议按照这两个数量的比值分配赌金。换句话讲，假设两人在接下来的游戏中旗鼓相当，赌金应该按照每个玩家在经历一系列对局后最终取胜的概率来分配。对概率的系统研究由此拉开了序幕。

这个问题能被概率的客观方法解决，但是帕斯卡考虑得更多。他提出了一个有关上帝是否存在的赌局。“上帝存在，或不存在，缘由无法回答。在无限远的彼岸掷一枚硬币，正面或者反面就要出现。你赌哪边？”

他提出，如果上帝存在，相信或者不相信带来的不同，就是在天堂获得无限的幸福与在地狱忍受无限的痛苦之间的区别；如果上帝不存在，相信或者不相信只会带来尘世生活中细小的差别。所以一个不可知论者应该强烈倾向于相信上帝存在。

在这个赌局中，“正面”或者“反面”出现的概率大小是具有个人色彩的选择，不能从对称性抑或计数证据中推导出。所以帕斯卡也是概率的主观方法的先驱者。

瑞士的伯努利家族

在17世纪和18世纪，来自巴塞尔的伯努利家族[1]的成员在数学(包括概率)领域取得了重要进展。家族内的竞争起到了鞭策作用：他们中的一个会提出难题，另一个就会回应，难题的提出者会说他发现了所谓的解决方案中的瑕疵，等等。

关于概率的游戏激发了许多对概率运作的早期关注。在这些游戏中，无论是掷色子、发牌，还是掷硬币，一些“试验”会在本质上相同的情况下被重复进行。之前提出过一个自然的问题：一个结果被观察到的概率和客观概率有什么关系？

雅各布·伯努利(Jacob Bernoulli)在其遗作《 猜度术》 (The Art of Conjecturing)中，用他的例子巧妙地进行说明，给出了一个答案。假设罐子中60%的球是白色的，其余的是黑色的，随机抽取一个球。伯努利证明，只要试验抽取至少25 550回，每一次试验中抽到白球的比例落在58% ~ 62%的范围**外部**，就会有至少1000次试验中抽到白球的比例落在这个范围**内部**。不规范地说就是：在多次抽取的条件下，我们观察到白球的频率会压倒性地倾向于接近它的客观概率。

类似的分析过程适用于任意能在相同条件下不限次数地重复的试验，一个试验的结果不会对其他试验结果产生影响。每一次试验中，某些特定的结果代表着成功，它们的客观概率是一个固定的值p。这个概念现在被称为**伯努利试验**(Bernoulli trials)。在p这个值附近取任意区间，你愿意它有多小就有多小(± 2%或± 0.1%，都无所谓)。然后给出你想要让成功的频率落在这个区间内部比落在其外部高多少(100倍还是100万倍，怎样都行)。伯努利的方法证明了只要试验重复足够多次，**任意**这样的要求都会被满足。观察到的频率会像你期望的那样尽可能地接近于客观概率，只要给出充足的数据。这个断言被称为**大数定律**(the Law of Large Numbers)。

在1975年，一个主要致力于促进概率和数理统计发展的国际学会被命名为“伯努利学会”，以向这个家族致敬。

亚伯拉罕·棣莫弗

亚伯拉罕·棣莫弗(Abraham de Moivre)以胡格诺派[2]难民的身份在英国定居，依靠国际象棋和他的概率知识谋生。艾萨克·牛顿(Isaac Newton)那时已经50多岁而且事务非常繁忙，为了岔开有关数学的咨询，他说：“去找棣莫弗吧，他比我对这些事情了解得更清楚。”棣莫弗的《机会的学说》(Doctrine of Chances)在1718年以英语出版，1738年的第二版包含了伯努利工作中的主要进展。了解他的成就要思考一个具体的问题：如果一个公正的色子被投掷1000次，我们能合理地预期数字6的产生频率与平均频率之间有多大偏差吗？

棣莫弗提出了一个对这类问题具有广泛应用的公式。他高超的洞察力表现在，他意识到数字6的实际数量与期待的平均数量之间的偏差，可以用投掷次数的**算术平方根**来进行最适当的描述。

如何夸大这个发现的重要程度都不为过。当你听说一个民意测验(opinion poll)中一个政党的支持率是40%，它经常会附加一个暗示，这只是一个估计，真实的支持率“非常可能”在一个范围中，比如38% ~ 42%。这样的区间宽度告诉你最初数字40%的精确度，而如果你想要更高的精确度，就需要更大的样本：这个平方项意味着要将精确度变为2倍，样本需要扩大4倍! 我们有一个“报复式”的收益递减法则——要达到原来的2倍效果，我们必须投入原来的4倍精力。

棣莫弗的方法可以用考察一个公正的硬币投掷20次时有多少次正面朝上来说明。基于所有的例如正正正反正……正反正反这样的，长度为20的序列都是等可能出现的，我们可以绘制出图1。其中垂直条的高度表示大约100万种序列中有多少个恰好包含0、1、2、……19、20

个正面。这些数字各自的客观概率就正比于这些高度。棣莫弗证明了经过这些竖条顶点的最佳拟合的光滑连续曲线非常接近于一个特别的形状，现在通常称之为**正态分布**(normal distribution)。

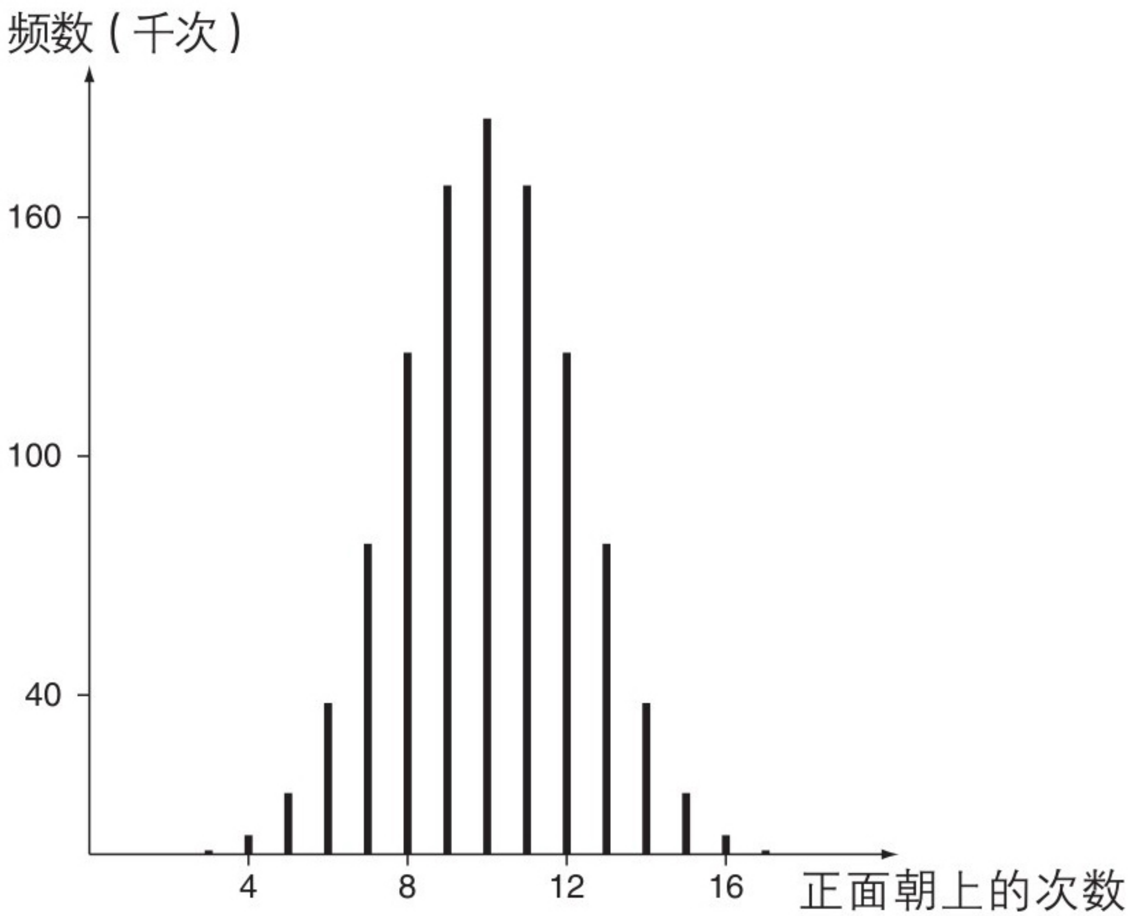


图1　20次投掷中正面朝上的相对频率

这种曲线会生成于所有的多次掷硬币过程中，并且还可以包括掷出正面的概率不等于1/2的情况。所有曲线之间有一个简单的关系，所以棣莫弗可以就一个基本的曲线制作一个简单的数表，并能在任何情况下使用。整体的成功频率在一个确定的限制范围中，现在就能够简单地获取对这样的事件的发生比例的估计——需要的仅仅是获胜的概率和将要进行的试验的次数。将一个公正的色子掷200次，你想知道数

字6出现次数在30~40间的可能性有多大吗？或者一个公正的硬币在100次投掷中掷出60次以上的正面的可能性有多大？没问题——棣莫弗有解决方案。

假设我们知道一群人死亡时的年龄，所有人都活到了至少第50个生日。棣莫弗的工作可以回答这样的问题："如果一个50岁的人在70岁之前死亡是更有可能的，我们能够观察到这些数目的各种变化的可能性有多大？"虽然这十分有用，但是它不能回答新兴的人寿保险业提出的关键问题："我们有多么确信一个50岁的人在他70岁之前即死亡是更有可能的？"

逆概率

托马斯·贝叶斯(Thomas Bayes)是一个在数学领域有建树的长老会牧师，他的思想现在比在其生前更受重视。他的《机遇问题的解法》(Essay Towards Solving a Problem in the Doctrine of Chances)在他死亡三年后的1764年出版，给出了初步处理主观概率的一般方法和从数据中推断概率的保险精算师问题的一个解决方法。这本书也包含了一个处理概率的重要工具，被称为**贝叶斯法则**(Bayes' Rule)。

为了举例说明这个法则，设想我们掷一个公正的色子两次。已知第一次掷色子点数是3，很容易地就能够得到总点数是8的概率，因为这个事件会在第二次投掷点数为5时发生。我们不假思索地就能给出解答为1/6。但是将问题调转一个方向，设问：给出总点数是8，第一次掷出3的概率是多少？答案远远不那么简单了，但是我们可以应用贝叶斯法则来得到结果。在掷色子的标准模型下这个概率为1/5。

对于刑事审判中处理证据的方法，**逆概率**(inverse probability)这个概念至关重要。假设在犯罪现场找到的指纹被鉴别

为属于一个已知的人——史密斯。如果史密斯是无罪的，发现这个证据的概率很可能是非常低的。但是法院判决的依据不是“已知史密斯是无罪的，发现这个证据的可能性有多大”而是“已知发现了这个证据，史密斯无罪的可能性有多大”。贝叶斯法则是获得答案唯一合理的方法。我们将会在后面的章节中看到这个法则是如何帮助我们做出正确决定的。

贝叶斯展示的洞察力在很多年中被忽略了，但是他的确指出了中心问题：如果在一系列的伯努利试验(例如掷色子)中，成功的概率是未知的，但是试验和成功的次数都分别是已知的，这个不可知的概率落在指定区间内的可能性有多大？而另一位极其优秀的数学家拉普拉斯的计算优于贝叶斯。

从1774年试探性的开始到1812年的理论综合体，拉普拉斯逐渐地完善着他的分析，并最终给出了解答贝叶斯问题的一系列明晰的公式。例如，利用巴黎男性和女性的出生人口数目，他得出结论，毫无疑问男性出生的概率高于女性——他估计这结论错误的概率是10^{-42}。

贝叶斯被安葬在伦敦的邦丘原野公墓(Cemetery of Bunhill Fields)，在皇家统计学会(the Royal Statistical Society)附近。其墓地曾经被修复过，来表达全世界统计学家对贝叶斯的敬意。

中心极限定理

将一些伯努利试验的结果写成由胜利(Success)和失败(Failure)组成的序列，例如FFFSFFFSSFSFF……现在将每个S用数字1代替，每个F用数字0代替，得到0 001 000 110 100……这表明了一个巧妙地理解这些试验中胜利的总数的方式：序列中的这些数字的**和**(同意吗？)。棣莫弗利用他所谓的正态分布曲线，给出了一个描述这个和的分布的良好近似方法。

对于一个巨大的数值序列，我们要考虑的可能只是其中随机变化的各个值的和。例如，负责垃圾处理的政府部门主要感兴趣的是整个城镇中的垃圾总量，而不是来自每个家庭的数量。当一位园丁播种红花菜豆时，他关心的不是每个豆荚的大小，而是总产量。一个赌场基于它的全部赢得的钱来评估其经济收益，不论个别赌徒的收益如何。将着眼的事物看作大量随机数据的和，这经常是卓有成效的。

拉普拉斯拓展了棣莫弗的工作以便能涉及像这样的情况。他建立了**中心极限定理**(Central Limit Theorem)，该定理说明了在很多情况下，大量随机数据的和是棣莫弗的正态分布的理想近似状态。我们不需要某个单独数据如何变化的细节，**整体数据**变化的模式会紧密地贴合这个正态法则。

为了利用这个想法，我们只需要两个数字：第一个是全体数据的平均值，第二个是一个简单地表示它的变化程度的数据。知道这两个数据，任何一个概率都能够从棣莫弗的表格中找到。

下面谈到卡尔·弗里德里希·高斯(Carl Friedrich Gauss)，他是和牛顿、阿基米德(Archimedes)并列的顶级数学天才。当时他正在研究如何处理观测恒星和行星位置时产生的误差。他提出，平均而言误差为0——观测中向左偏差和向右偏差是等可能的——**并且**误差大小遵循这个正态分布，他因为这个方法在数学上很简洁而使用它。但是拉普拉斯看到高斯的书时，将这个结果引用到了自己的书中，同时提出，因为一次观测中的**全部**误差以许多随机因素堆积总和的形式出现，这样误差**应该**遵循正态分布法则。高斯蹩脚的理由“数学上的便利”被拉普拉斯更加有说服力的“数学理论表明……”所替代。

“正态”这个术语应用在这个分布上是不恰当的。它暗示我们应该预期我们遇到的任何数据都遵循这个形式，但是远非如此。为了避免这种暗示，并且为了纪念一个伟大的人，我们将会转而使用另一个

术语**高斯分布**(Gaussian Distribution)。如果你可以说服自己，你关注的值可以貌似可信地被当作大量随机的较小数字的和，这些较小的数字的来源都是不相关的，那么这个中心极限定理告诉我们可以预期这个值遵循高斯分布。

观测中的误差真的遵循这个规律吗？亨利·庞加莱(Henri Poincaré)——对数学各分支具有全面知识的最后一位数学家——说：“人人都相信它，因为数学家误以为这是观测中的事实，而观测者认为这是个数学原理。”

西莫恩·德尼·泊松

西莫恩·德尼·泊松(Siméon Denis Poisson)出名是因为一个含有他名字的**分布**——概率在一个平均值周围变化的方式。在物理学家欧内斯特·卢瑟福(Ernest Rutherford)及其同事的工作中——计算7.5秒长的时间间隔内有多少个α 粒子从放射源中发射——相关的例子出现了。这个数字从0到十几不等，平均值稍小于4。图2展示了两个经典的实验，说明(在这些情况中)有四五个粒子发射。卢瑟福认为这些发射事件都是随机的。

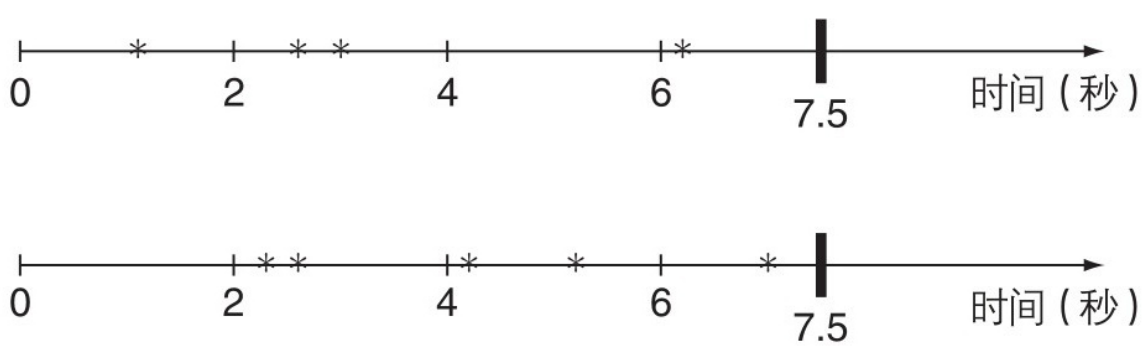


图2　α 粒子的发射时间

将7.5秒切成极多个极小的时间间隔，小到我们可以忽略其间发生一次以上发射的概率。除了几个发生了一次发射，其他所有的小间隔都没有出现发射这一事件。在各个小间隔之中，将一次发射事件看作游戏胜利，所以粒子发射的总数就是胜利的次数——又是伯努利试验。

极小间隔胜利的概率实际上和其长度成比例，所以随着单个间隔长度的缩小，间隔的数目增多，每一个间隔发生胜利事件的概率减小。泊松计算出了小区间长度缩小至0的过程中，发生0、1、2……次发射的全部的确切概率。

这个**泊松分布**(Poisson Distribution)就经常出现在我们计算事件“随机”发生概率的时候，至少是以一种良好近似的方式。它恰当地描述了卢瑟福的实验数据；它适用于第二次世界大战中投掷在伦敦南部不同地区的炸弹数量；它看起来是一本书中每1000个字中错印字数的有效模型。如果你同时从两个洗好了的牌堆发牌，正面朝上，平均而言你会发现发牌恰好有一次是一样的，但是实际的匹配次数会非常接近于泊松分布。有一个可怕的例子，在一项长达20年，涉及几代普鲁士特兵团学员的追踪调查里，被自己的马踢死的军官数量也满足这个分布。

所有这些例子都符合一个相同的模式：大量的机会，每个机会中事件发生的概率很小。每当你正在研究的现象符合这种模式，泊松分布就很可能对它有用。

俄罗斯学派

一个数学定理具有这样的特征：如果某一个假设是成立的，那么一个预期的结论就会产生。我们主要的兴趣在于应用这个预期的结论，所以所需的假设最好不要太复杂。有时候预期的结论只存在于非

常具有限制性的假设的情况下，或者极其难被证明：之后的研究者也许会找到使用相同假设的更简单的方法，或者在较少的限制性条件下得到相同的结论。最好的情况是结论在非常宽松的假设成立时，能被简短而优美地证明。巴夫尼提·切比雪夫(Pafnuty Chebychev)的工作给出了这种理想情形的良好案例。

切比雪夫展示了如何在更广泛的情况中应用大数定律。最初的大数定律和伯努利试验有关，它描述了在一系列试验中事件发生次数的比例能够多么合适地用于估计事件发生的概率。如果我们想估计入伍士兵的平均身高，或者一个家庭一周的消费，我们似乎可以很明显地从相关人员中抽取合适的样本。但是这种估计有多合适呢？切比雪夫的工作给出了误差足够小以使得估计可靠的概率。

很多统计结果都是这些想法的应用。

切比雪夫最知名的学生是安德雷·马尔可夫(Andrey Markov)，马尔可夫的教学启发了又一代极有天赋的俄国人。马尔可夫将他的想法应用于诗歌和文学作品。在将亚历山大·普希金(Aleksandr Pushkin)的《叶甫盖尼·奥涅金》(Eugene Onegin)中的元音字母(vowel)和辅音字母(consonant)分别替换为字母v和c之后，马尔可夫得到了一个只含有这两个字母的序列。在原始的基里尔字母[3]中，元音字母占有43%的文字比例。在一个元音字母之后，另一个元音字母出现的频率是13%，而在辅音字母之后，元音字母出现的频率是66%。在预测一个字母之后的字母是元音还是辅音的过程中，他发现，已知当前的字母时可以忽略它前面所有的字母，因为它们基本不构成影响。

这个“可以忽略”的特性广泛存在。有一些例子：赌徒手中的一系列连续的赌金数额；特拉维夫(Tel Aviv)每天的天气(干或者湿)；在每一个顾客离开时队列的长度；连续世代的基因组成；两个相连的

容器中的气体扩散过程。如果知道序列中的前一个值，要预测随机变化序列的下一个值的时候，我们都可以忽略更前面的那些值，那么这个序列被称为具有**马尔可夫性质**(Markov property)。描述这种序列的理论已经很好地发展起来了，这些理论也是许多概率成功应用的基础。

马尔可夫在政治上非常活跃，对数学史也知之甚详。1913年，俄国政府组织了罗曼诺夫改革300周年的庆祝活动，马尔可夫相应地开展了对伯努利发现第一个大数定律200周年的纪念活动。

这里我偏离一下主题，讨论一下在20世纪早期，法国人埃米尔·博雷尔(émile Borel)的工作。回想伯努利试验中的大数定律：在大量试验中，实际事件发生的频率有极大的特别接近于事件发生概率的可能性。但是这仍然留下了这样的可能：在无限次的试验中，实际事件发生的频率极其偶尔地会落到事件发生频率附近任意给定的公差带之外。但是博雷尔的工作完全消灭了这种概念难题。任意给定公差带，一定会有一个时刻(我们不知道什么时刻，但是的确有那么一个时刻)，在这个时刻之后，实际事件发生的频率就会永久地停留在公差带**内部**。这被称为**强大数定律**(the Strong Law of Large Numbers)。

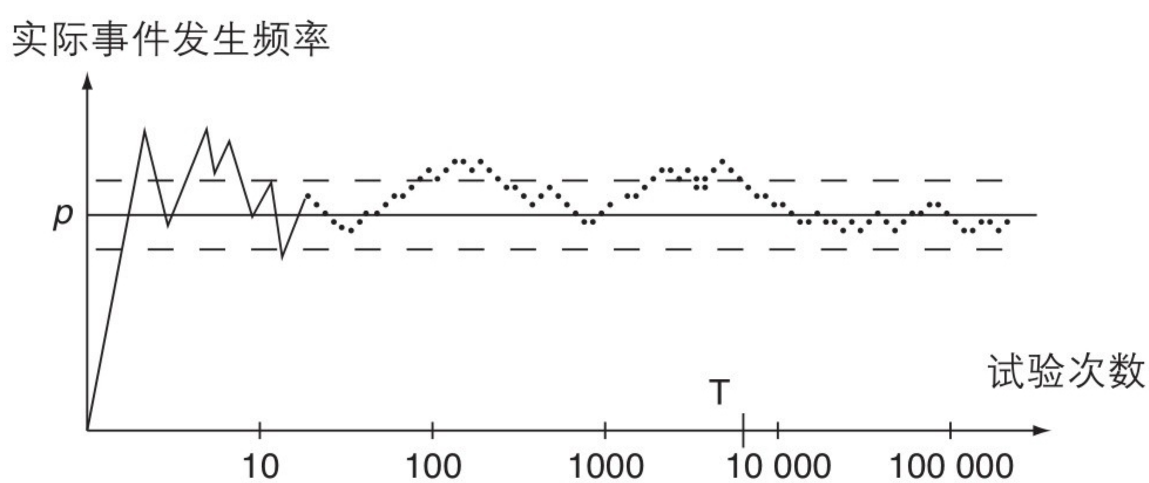


图3　强大数定律图示。p是事件发生的概率，虚线表示公差带。在T次试验之后，实际的事件发生的频率永久地停留在公差带内部

这个强大数定律也会延伸到更广阔的情境中。我们可以将大数定律的含义浓缩到一个不正式的习语中：

> 从长远来看，平均统领一切。

在1924年，亚历山大·辛钦(Alexander Khinchin)发表了命名极佳的《重对数律》(Law of the Iterated Logarithm)。就像伯努利和拉普拉斯的早期工作一样，这个理论应用于一个以和的形式出现的随机数量时，它可以给出更加精确的有关这个和会多么接近其平均值的信息。

在大约300年中，概率工作的前沿进展都是来自一些特别的方法。然而到了1933年，杰出的苏联科学家安德雷·柯尔莫哥洛夫(Andrey Kolmogorov)使用了近期发展出的**测度论**(measure theory)中的理念，将概率科学确立在了令人满意的逻辑框架中。所有已知的理论都可以重新容纳进柯尔莫哥洛夫的设定中，并能给出精确程度作为后续计算过程的催化剂。

柯尔莫哥洛夫和辛钦还有他们的学生鲍里斯·格涅坚科(Boris Gnedenko)一起极大地拓展了拉普拉斯关于随机数量和的工作。提高纺织业和其他制造业中机器的可靠性的方法，生产线上产品的质量控制，还有解决拥堵导致的问题都是他们研究的动机。

柯尔莫哥洛夫是一个卓越的研究者和教育者。他在1987年去世，当时的苏联总书记米哈伊尔·戈尔巴乔夫(Mikhail Gorbachev)还调整了自己的工作日程，以便能够出席葬礼。

更近的近代

战争经常会驱动科学发展。1939—1945年的世界冲突促进了运筹学的发展，其中许多成就都起源于对概率论理念的合理利用。为了使一艘补给船不被敌军潜艇击沉的概率最大，一系列数据的组合和计算给出了舰队比单艘船更好，大舰队比小舰队更好的结论。当这个结论被付诸实施，损失就显著地降低了。布莱切利园(Bletchley Park)的密码解读工作现在大概已经广为人知了。然而我们经常忽视贝叶斯公式在确定恩尼格玛密码机(Enigma machines)卷轴最有希望的布置方式中的应用。

在1950年，威廉·费勒(William Feller)出版了一本介绍概率的著作，并在1957年和1968年再版。这本书是我心目中有史以来最好的非虚构著作。凭借着直觉和严密论证的结合，这本书直接地或者间接地引发了人们对这门学科的巨大兴趣。随后，乔·杜布(Joe Doob)使用术语“**鞅**”(martingale，这个词原本指每次损失后将赌金加倍的投注策略)来描述那些在未来某时刻的平均值与现在的值(大致上)相等的随机量。他揭示了鞅的主要性质并给出了一些密切相关的概念：这些工作普遍有用，因为它指出许多有实用性的随机量都会包含在这个理论研究的范畴中。之后我们会举例说明概率这个概念是如何在一系列领域中被有效地应用的。

许多专攻概率的学术期刊已经发行，其中一些还产生了分支期刊，从来没有听说它们缺少可以发表的材料。现代计算机的算力已经转变了计算概率的模式：它们的运算速度和存储容量极大地拓展了可解决问题的范围。在早期，许多概率问题只受一个因素影响，比如说时间或者距离，人工的精确计算总是可能的；而现在，那些概率随时间、空间的三维和其他因素的影响而变化的复杂问题也已经被成功解决了。

即便如此，使沟通交流更便利才是计算机对概率论发展最巨大的影响。TeX[4]语言已经成为数学和许多科学写作的标准语言框架。研究者在互联网上发布他们的想法和观点，学术文章可以在家中或者办公室中通过万维网(the World Wide Web)轻易地取得。

[1] 伯努利家族(The Swiss Family Bernoulli)来自瑞士的一个商人和学者家族，有很多艺术家和科学家出自其中。

[2] 胡格诺派(Huguenot)， 16世纪至17世纪法国基督新教归正宗的一支教派，1685年被法王路易十四宣布为非法。

[3] 基里尔字母(Cyrillic alphabet)又称西里尔字母、斯拉夫字母，是在使用斯拉夫语族语言的大多数民族中通行的字母书写系统。

[4] 正式标志为TEX，中间的E有点下沉，但在无格式纯文本中写为TeX。

04 概率试验 Chance Experiments

对于以概率作为结果的任意试验——买彩票、投注赛马、相亲、接受医学治疗，我们用**分布**这个词来详细说明其所有可能的结果，以及与它们相关的概率。我们讨论泊松分析——大量重复试验中多少稀有事件会发生——的时候提到过这个词。

“分布”是分析概率试验中结果变化范围的中心概念。坦率地说，我们需要知道可能的结果的范围。为了给出这些结果概率的合理数值，我们必须讲清楚我们的假设，并且期望它们对于我们想要考察的试验是合适的。

离散分布

首先，我们来看看那些可能的结果能够被写成一个列表的情况，每个结果都有它们自己的概率。术语**离散分布**(discrete distribution)适用于这种情况。

最简单的情况就是我们认为结果具有相同可能性时计算结果的数量。这里使用**均匀分布**(uniform distribution)这个术语，因为总体的概率均匀地分散在各个结果上。许多试验都被认为满足均匀分布——轮盘赌、掷色子、扑克牌、选择彩票中的中奖号码等。精确的计数给出了合适的答案。

术语“伯努利试验”描述了一系列发生概率均为常数的独立试验。在伯努利试验次数固定的情况下，有一个简单的公式叫作**二项分布**(binomial distribution)，分别给出了事件发生恰好0、1、2……次的概率。这个公式只依赖于试验的次数和事件发生的概率。当你依

次浏览这些结果的时候，它们的概率先是升高到一个最大值，然后逐渐降至0。泊松分布也遵循这个模式。

我们能计算20次掷色子中数字6出现次数的二项分布；或者一个学生对30道多选题中的5个选择随机作答的时候，蒙对个数的二项分布。但是我们**不能**预测一个桥牌选手的13张手牌里梅花张数的概率：虽然每一张单独的卡片都有1/4的概率是梅花，但是连续的牌不是独立的，因为下一张牌是梅花的概率会被所有前面的结果影响。

永远要留意(通常是小号字)附属细则。使用二项分布需要3个条件：固定试验次数，每个事件与其他事件相互独立，并且事件发生的概率是常数。

在一系列伯努利试验中，事件首次发生的时候经过了5轮试验的概率是多少？这种情况发生的唯一方式就是前4次试验中事件都未发生，随后1次试验中事件发生；因为所有的试验都是独立的，问题的答案就是将这些结果分别的概率乘在一起，给出了一个令人愉快的简介表达式，这就是所谓的**几何分布**(geometric distribution)。

事件首次发生需要恰好1、2、3……次试验的概率稳定地下降。下一个概率值等于将现在的概率值乘以一次事件未发生的概率，一个小于单位1的固定值，每次都是这样。因此，无论事件发生的概率是多大，事件首次发生时经过的试验次数的最可能的值就是1。

假设在板球比赛中，连续的击球构成了伯努利试验。一位投球手，将事件发生理解为他投球成功，他可以乐观地想：他开始投球的时候，下一次投球成功最可能的就是这一次；相反，一个具有相同视角的击球手就得听天由命地接受他这一局最有可能的持续时间就是他面对这一个球的时间。(就算是最好的击球手，记录表明他们最有可能的总得分总是0!)

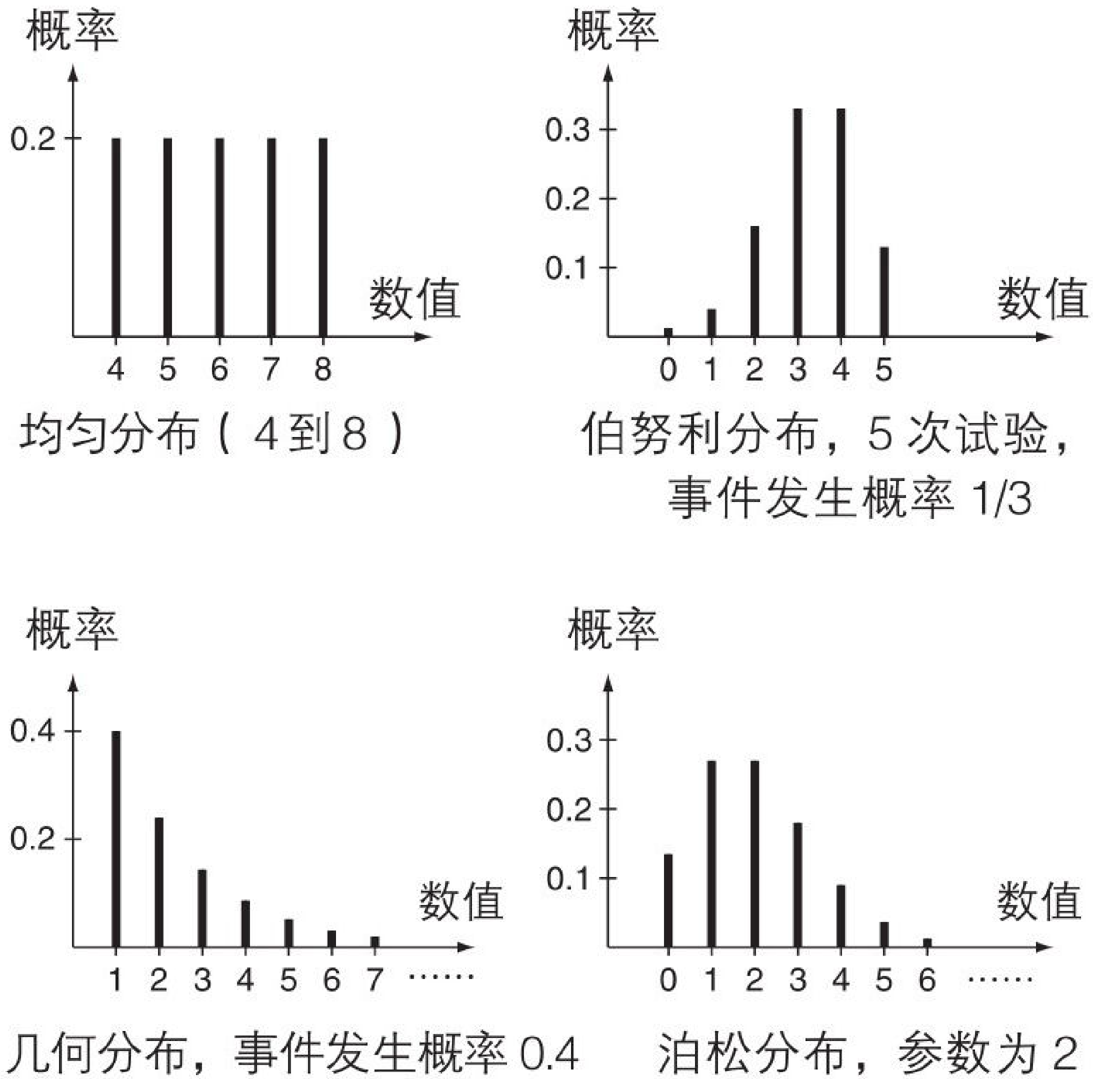


图4　一些常见的离散分布

图4举例说明了一些常见的离散分布。对于每一个可能的数值，竖线的高度给出了它的概率，并且这些高度的和总是单位1。

连续分布

我们如何拓展古典的概率观点来解决在一个长度为80cm的木棍上随机选取一个点的试验？可能的结果组成一个**连续统**(continuum)，

而不只是一个列表。

“随机”意味着所有单独的点都具有相同的概率值。但如果这个相等的值超过了0，那么，在取了足够多的点之后，它们的总概率就会超过单位1，这是不可能的。每个单独的点的概率一定是0，我们也不能使用像图4一样的图片了。我们需要将概率、片段或者区间相关联，而不是将概率和单独的点相关联。

为了对80 cm的木棍的每一部分一视同仁，所有具有相同长度的片段一定有相同的概率。想象一下将木棍砍成8个相等的片段：按照定义，一个“随机的”点落在每个片段上的概率一定相同，举例来说，落在20~30 cm的片段上一定具有1/8的概率。

图5a展示了下一步操作，这可以用口头禅“面积表示概率”表述。标注了h的水平线的高度是设定好的，这条线下阴影部分的面积是单位1，这呈现了一个事实，我们可以百分百地确定随机点落在区间0~80 cm中的某处。接着图5b展示了如何确定随机点落在32~52 cm的片段上的概率，只需要计算对应的阴影面积即可。简单地说，这个概率是1/4。

要得出随机选择的点落在木棍两端10 cm内，或者中间20 cm内的概率，我们就可以使用图5c，并且依据加法定理，要求的概率是三个阴影面积的和，也就是1/2。

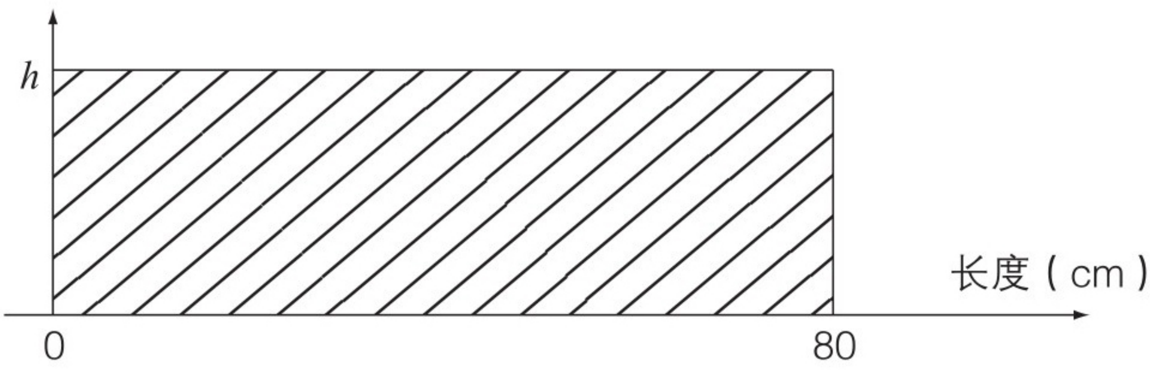


图5a　阴影面积是单位1

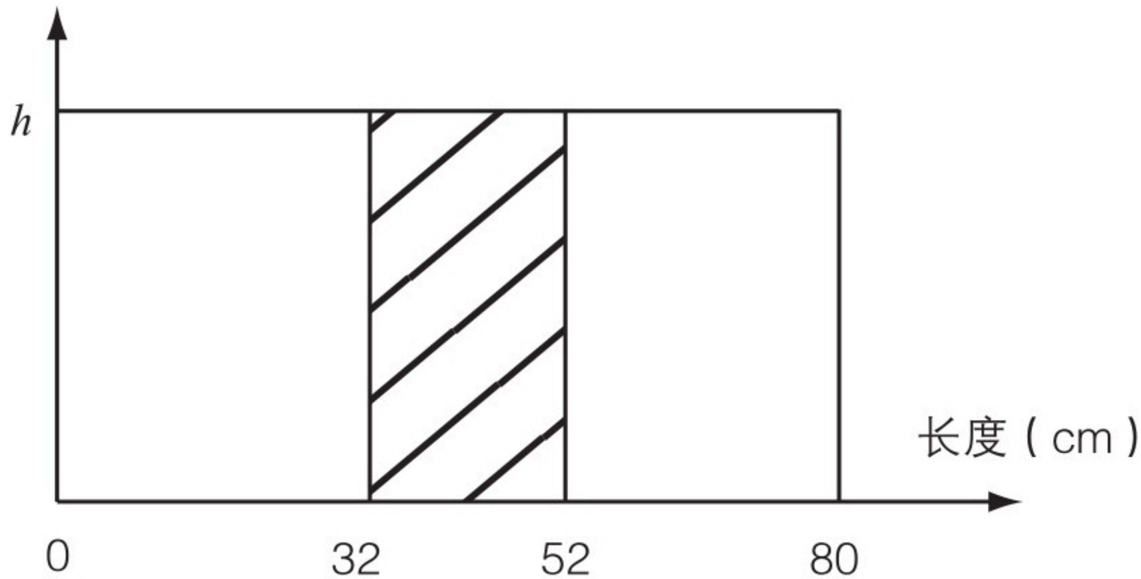


图5b　落在32~52cm之间的概率是1/4

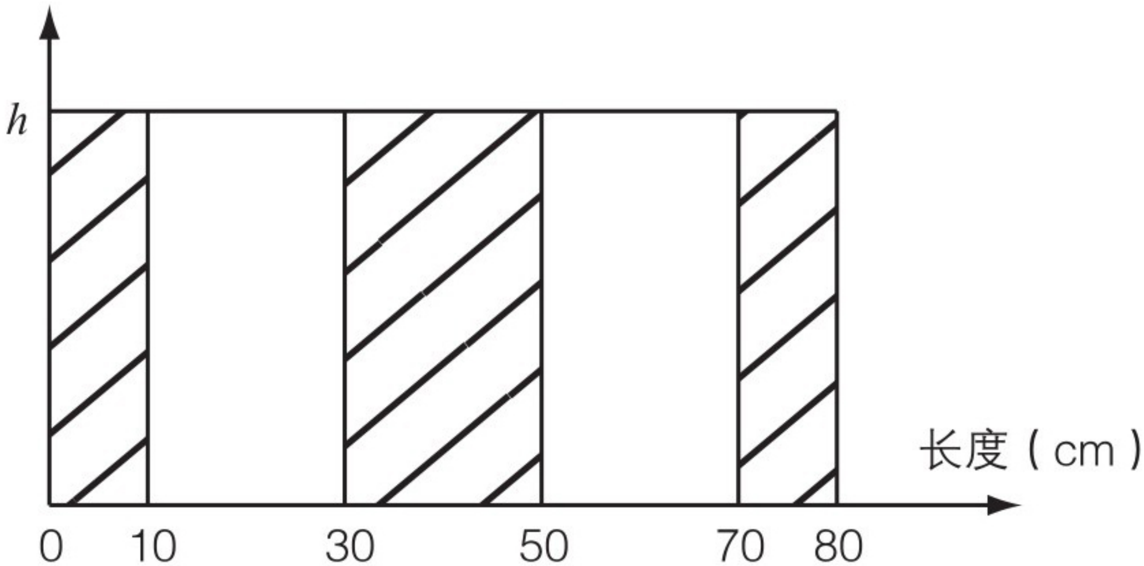


图5c　见正文

图6展示了对结果取连续值的另一些情况下相似的解决方式，例如一段特定的高速公路上下一次事故发生需要的时间。我们会在下面论证展示图片上的曲线在这种情况下是合理的，但核心观点是图线的尺度是特意选择好的，以至于标注了“时间”的直线以上，和以点E为起始端点的曲线以下的总面积是单位1，因为我们可以百分百确定我们考察的这段时间一定取非负值。

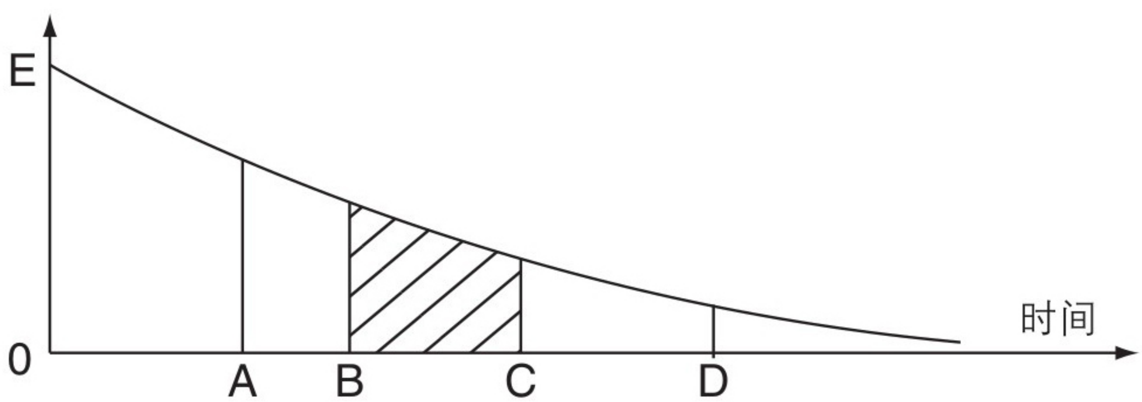


图6　连续分布

时间至少是B但不大于C的概率就是阴影的面积。我们可以用类似的方式得到考察的时间落在任意给定区间内的概率，还能像之前一样根据加法定理，得到落在更复杂区间内的概率。

一个能按照这种方式生成概率的曲线被称为**概率密度**(probability density)。已知面积的计算方式是“长乘以宽”，任何直线的宽度都是0。因此图6中在点A或者点D的竖直线的“面积”都是0，所以这两个单独的点具有0值的概率，就像之前提过的那样。但是点A的密度曲线比点D高，所以点A**附近**的值比点D附近的值更可能。简单地说，图片表明具有或高或低的概率的区域。在这里我们使用**连续分布**(continuous distribution)这个术语。

在所有这些试验中，因为单独一个点具有的概率值为0，我们可以稍微草率一点：无论一个区间包括了两个端点或一个端点，抑或都不包括，结果的概率都是一样的。

为了限定一个概率密度，一条曲线一定必须具有两个特性：不能取负值，在曲线下的全部面积必须是单位1。这些保证了对概率的所有计算能得出合理的结果。

许多概率密度函数出现得足够频繁以至于可以被赋予名称。对于从给定的一个区间内选取随机点，密度函数在这个区间内完全平直，就像图5中的一样：简单地说，所有相同长度的片段具有相同的概率。再一次，我们叫它**均匀分布**。

假设我们对一些特定事件在多长时间后发生感兴趣。例如，^{210}Pb是一种铅的不稳定同位素，“它的半衰期是22年”这个断言被印在物理教材上。它的意思是，如果我们有一块这种物质，22年后只有原来的一半保持原样，其余的都通过辐射衰变成其他物质了。

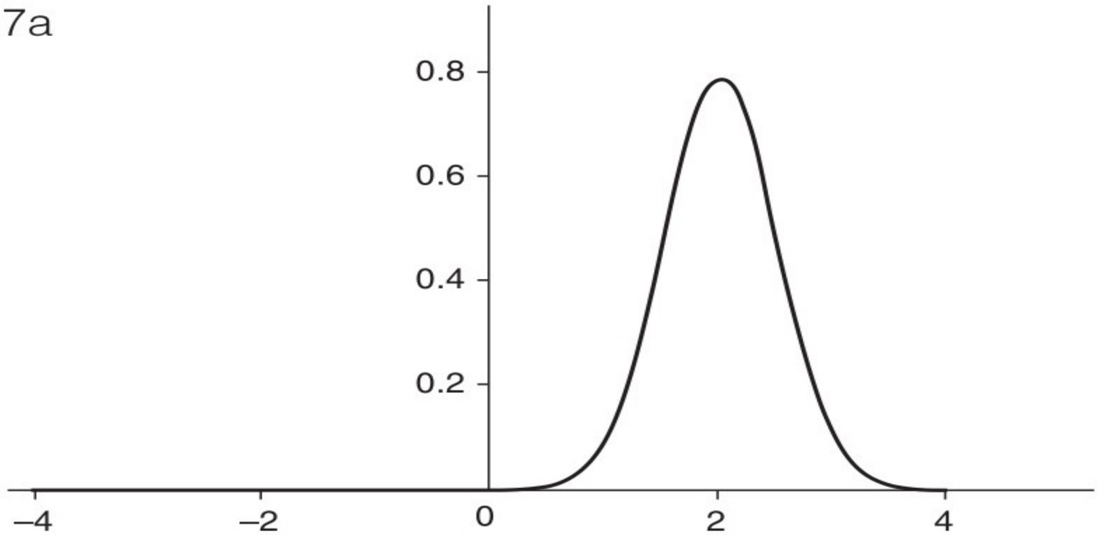
这块物质由巨量的原子组成，所有这些原子的行为都是独立的。如果关注单个原子，它通过放出一个粒子而衰变。我们不知道什么时候这个过程会发生，但是因为在22年内这块物质中的一半的原子都衰变了，所以这个**特定的原子**在这个时间段内发生衰变的概率是50%。假设它在5年后还没有发生衰变：这时，它就是剩余的^{210}Pb块中的一个原子，所以它在未来22年衰变的概率也是50%，并且如果它在接下来的3年中没有发生衰变，情况也一样，以此类推。

一个给定的原子的衰变时间只有遵循所谓的**指数分布**(exponential distribution)的时候，上述情况才能发生，它的概率密度的一般图形展示在图6中，曲线的高度按照确定的比率下降。类似的场景也应用在交通事故中：过去的一周内没有事故发生，那似乎不可能对未来的事故概率产生影响，所以我们预期交通事故下一次发生的时间也遵循指数分布。

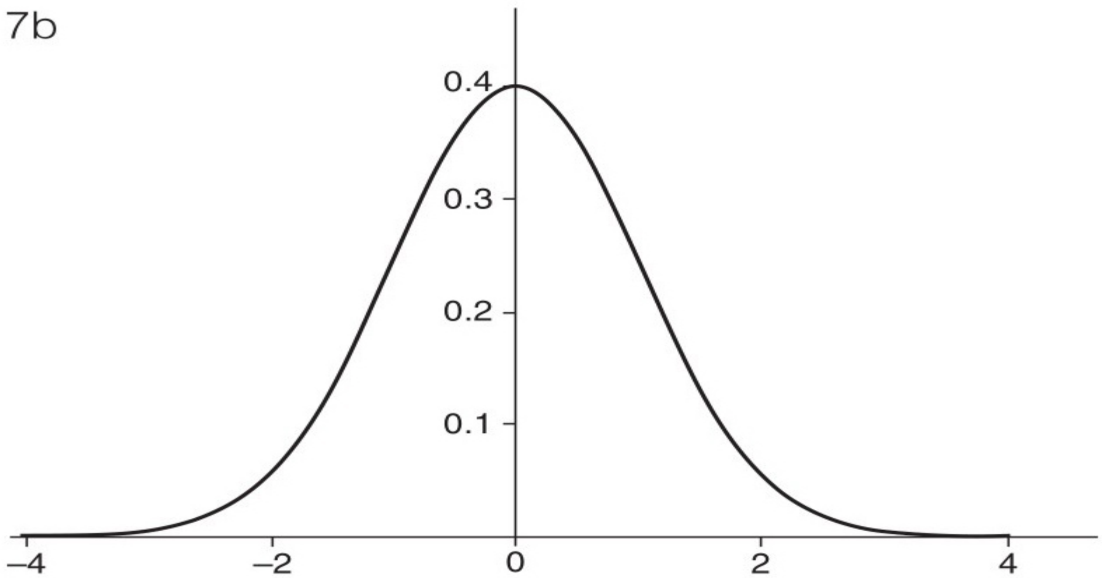
这个分布和泊松分布密切相关。只要事件本质上是随机发生的——暴风雨中的闪电、细胞复制中的自发突变、顾客来到邮局——在固定的时间段内这样的事件发生的**数目**倾向于遵循泊松分布，那么一对事件之间的**等待时间**的概率分布就具有这种指数形式。

最重要的连续分布是我们已经提到过的**高斯分布**。就像图7展示的那样，这个分布家族中的成员关于单独的一个峰对称，并且在两边快速下降，然而永远不会达到0。两个数字就可以告诉我们任意一个分布实例在这个家族中的归属：一个数字表示峰的位置，另一个数字描述散布程度——较小的散布值导致像图7a那样的高且窄的图形，较大的散布值给出像图7c那样的矮且宽的图形。这个家族成员的任何位置的概率都可以借由这两个数字来与图7b的分布相关联而得到，这个分布的峰在0，标准散布值为单位1。棣莫弗创制了对应数表之后，这些对应关系就很容易得到了。

7a



7b



7c

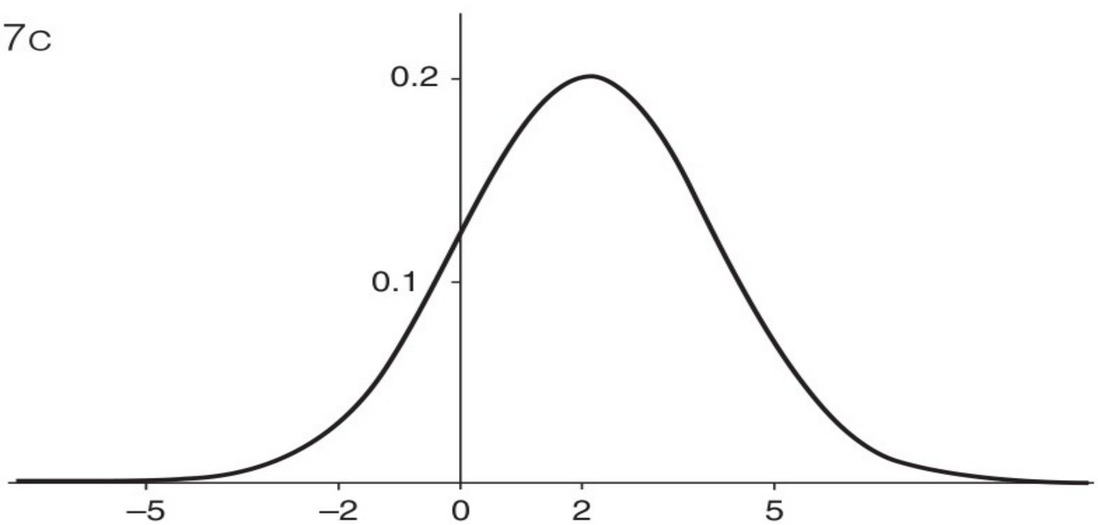


图7　高斯分布

一个问题的再解决

你也许已经注意到了一个问题。已知期望的结果组成的集合是有限的，或者是一个像{1， 2， 3，…}这样的无尽的集合，那么即使这个集合中的一些成员的概率是0，任何概率是0的事件也不会发生。然而，对于**连续分布**，即使每个单独的点的概率是0，它们其中的一个在试验进行过程中也是**会发生**的! 我们不再能够认为“不会发生”与“概率为0”具有相同的意义。

为了解决这个问题，我们来考虑从装有100万块完全相同的大理石的盒子中随机选取一块。只有在提前猜对了结果的情况下，我们才会感到惊讶，因为猜对的概率只有一百万分之一。但是，无论抽中了哪块大理石，虽然的确出现了某个概率只有百万分之一的结果，但我们也不会感到惊讶。

把盒子做大一点——10亿块或100 000亿块大理石——实际产生结果对应的概率可以无限接近0——但是它的确发生了。这与在一条连续的线上选取一个点的过程并没太大区别：对于任意的点来说，它的概率是0，但是它们其中的一个的确将会发生。

我们接下来开始说明，在一个可重复试验中，如果猜对结果的概率是1/6，我们可以期望按顺序进行的6次试验中有1次猜对。将事件发生的概率除以100万，我们预期等待正确结果出现的次数就被乘上了100万。具有极小概率的结果的确会发生，但是越来越罕见。

如果概率下降到0，我们可以预期要等待比任意有限长都要长的时间——那它就是不会发生! 在**提前指定**的情况下，认为任何概率为0的

事件都不会发生是合理的。

平均值

已知一个概率试验中结果的分布，我们就可以计算我们想要的任何概率。但是有些时候，所有的这些细节都成了障碍——只见树木，不见森林[1]：所以我们想要提取出分布的主要特征。

举例说明，假设可能出现的结果只有2、3和7，分别对应概率60%、10%、30%。我们预期在100次重复试验中，2这个值会出现大约60次，3出现大约10次，7为剩下的30次。所有这些数值的和是120 + 30 + 210 = 360，所以所有这100个结果的平均值是360/100 = 3.6。这个值就是数值2、3和7的**加权和**(weighted sum)，权重就是它们的概率。

无论我们有什么样的分布，相似的计算都会得出大量重复试验结果的集中趋势。“集中趋势”是一个宽泛的词，对于这类计算的结果，我们更喜欢使用**平均值**(mean)这个术语。可能有一些捷径：如果值在一个范围内均匀地分布，平均值就在两个端点的正中间；在一系列伯努利试验中事件发生的次数的平均值就是试验次数和事件发生概率相乘。

掷一个公正的色子，得到4的概率是1/6。所以在600次投掷过程中，我们应该可以得到大约100个4：简单计算表明，连续出现的4之间的平均间隔是6。大小为1/6的概率导致平均间隔是6，这不是一个巧合。任何间隔的长度就是下一个事件发生的等待长度，所以在一系列伯努利试验中，我们就有了令人愉快的结果：

> 等待一个事件发生所需的平均时间是事件发生的概率的倒数。

在连续分布中，想法是类似的，但是加权和是由一种名为**积分**(integration)的数学方法来得出的。对于高斯分布，峰处就是平均值。按照整体特定的频率发生的随机事件发生的平均时间是一个指数分布：平均时间就是频率的倒数，这并不奇怪。

除了“平均”和“平均值”，术语“期望”和“期望值”也会被使用。掷一个公正的硬币12次，“期望”正面朝上的数目为6；掷一个公正的普通色子，“期望”得到的分数是3.5。当然正是因为一次投掷中反面朝上的期待值是0.5，我们实际上不能**期望**得到一半的反面朝上! 文字都很是奇妙。

平均值非常友好：和的平均值通常是平均值的和，无论不同的和是不是独立出现的。大数定律告诉我们，从长远看，平均值占据主导：如果你买一张彩票花费1英镑，其中一半的钱都进入奖池中，那么，无论奖金的分配结构如何，你的**平均**收益都是50便士，从(非常)长远看，这就是你能得到的。

离散程度

用一种简洁的方式描述一个分布的离散程度通常是有用的。我们可以计算每一个值和平均值的差值，然后得到这些差值的(适当地加权的)平均值。但是，就像所有计算上努力展示的那样，这种方法是不成功的：负的差值不可避免地抵消了正的差值，最终结果总是0。

但是无论一个差值是正是负，我们都可以在将它平方之后得到一个正的值。所以我们可以通过将这些平方值加权来得到离散程度。得到的这个值就叫作**方差**(variance)。如果分布集中在平均值的附近，那么方差就会比较小；当有合理的原因使一些值距离平均值比较远的时候，方差就会变得很大。

当考虑以美元计收入的分布的时候，平方值的单位就是“平方美元”，不管它究竟是什么意思。将方差取算术平方根就可以得到原始计量中的单位，这就得到了**标准差**(standard deviation)。

平均值和标准差合在一起，经常能给我们理解一个概率分布的快速且有用的方式。在高斯分布中，仅用这两个数字就足够计算所有的概率! 就像点金石一样，当分布为高斯分布的时候，在大约68%的试验中，结果在平均值周围1个标准差范围内；在超过95%的试验中，结果在平均值周围2个标准差范围内；而400次中大约有一次结果是在3个标准差之外。

这些数字就是在第1章中给出的，我们能够有理由地期望事件发生的概率和事件发生的实际频率有多么接近的相关参考的基础：关键就是中心极限定理，它说明了作为大量随机成分的和而出现的数量预期接近遵循高斯分布。

在图7中，展示了3个高斯分布的概率密度函数，这几张图的平均值分别是2、0和2，标准差分别是1/2、1和2。

但是注意：虽然和的平均值总是平均值的和，但是对方差和标准差来说可不是这样。如果和的组成部分恰好是独立的——比如说在拉斯维加斯一家赌场7天分别的收益——那么和的方差就的确是分别的方差的和，否则就会偏高或者偏低。直接将标准差相加几乎不会给出任何有意义的结果。

极端值分布

在概率的某些应用场景中，我们关注的重点在于随机数量的最大值或者最小值。例如，线或者电缆的强度依赖于最弱的纤维；洪水防护设施要考虑的是下一个一百年中预期发生的最大规模的洪水；**生存**

分析(survival analysis)这个学科调查一段给定时间后的剩余人口。极端事件可能很少发生，但是当它们发生了的时候，结果就变得很重要了。

最简单的看似可信的模型假定存在一些独立的随机变量，每个都分别遵循一个特定的分布。例如每一年中，对一家保险公司的索赔。对接下来的50年中它可能会收到的最大的总索赔额有多大，保险公司有一个经历了漫长的数学推导的可用结果：无论在每一年中索赔额如何变化，在很大的年代跨度中，最大索赔额一共只有三个可能的种类，它们被称为极端值分布，具体的名字是弗雷歇(Fréchet)、冈贝尔(Gumbel)、韦布尔(Weibull)。有一个合理的数学原理，如果有一个关于最大值的理论，就一定有一个相对应的关于最小值的结果。所以如果感兴趣的东西是最小值，也存在相似的结论。

能够对这三种分布进行一些限制是非常有帮助的。通过估计极端值的平均值和方差，从三种分布中选择一种最接近于实际数据的，就能计算分布中的另一些概率的合理估计，比如真实情况中极端和破坏性事件的概率。

[1] 原文为“we can t see the wood for the trees”。

05 理解概率 Maki ng Sense of Probabi l ities

这里我会介绍在我们面对不确定性问题的时候，概率的观点是如何起作用的，还会描述一些误解了概率的情况。

赔率？

回忆之前我们说过，概率可以用赔率(odds)的形式来描述，反之亦然：概率是1/5和赔率是4比1是相同的。不幸的是，“赔率”这个术语也已经被赌徒们篡改成了一个相当不同的意思——如果你选择的马胜出了，你获得的赌资。所以当你听说“海都之星(Sea The Stars)”在2009年德比大赛(Derby)[1]中以11比4的赔率胜出，这只是说明了对于押在这匹马上的每4英镑赌注，它如果赢了收益就是11英镑。“11比4”没有与胜利概率之间的必然联系。这个比例只依赖于赌徒对于马胜出概率的主观评定。对于像“11比4”这类的表达，术语“赔付金额(pay out price)”是一个更准确的语言使用方式，但遗憾的是，我们不得不接受在赌博情境中对于“赔率”一词的滥用。

如果两边都没有经济收益上的优势的话，赔付金额就被称为“**公平的**”。也就是说，赌注的平均值是0。从洗好的牌堆中正确地抽出指定花色牌的公平赔付金额是3比1，这也就是抽中的赔率。

因为商业赌博不会在没有庄家收益(house advantage)的时候运行，就是说，商业赌博都是不公平的。对一家英国赌场的轮盘赌来说，所有的37种结果都是等可能的，赌其中一个数字的赔付比例只有35比1，而不是公平的36比1。所以一次37英镑的赌局回报的平均值只有36英镑，这就得出庄家收益率——对于任何赌局庄家都预期会赢——为1/37，大约2.7%。

这个优势对大多数你能想象得到的轮盘赌局都是一样的：无论你是赌一对数字，还是3个、4个、6个、12个，对于你下注的每37英镑，你的平均回报总是36英镑。在拉斯维加斯，标准的庄家收益还会更大一些，因为一个多加的格子，双零(double zero)会给出38种结果，但是赔付金额和在英国时相同。38美元的平均收益通常是36美元，庄家收益率是2/38，或者说5.3%。

赌马的另一种形式是通过同注分彩系统。这里，押在所有的赛马上的赌注被放入彩池中，以固定比例——常见的是80%——被押了获胜马匹的人按照他们下注的比例分享。无论哪匹马获胜，分彩收益率就是20%。

赛马中，赌注登记人(bookies)的收益的大小取决于哪匹马获胜了。虽然赌博者在单次比赛中都既可能收益也可能损失，但最近的数据给出了一个令人清醒的事实：在赔付金额为6比4的时候，赌博者预期会损失他们赌本的10%；赔付金额在5比1的时候，预期损失大约13%；在10比1的时候，平均损失超过了23%；如果你投机地定下50比1的赔率，预期你会损失2/3的钱。

这个现象被称为**热门冷门偏差**(favourite-longshot bias)。赌博者在更受欢迎的马匹上输钱会比高赔付金额的输钱**更慢**。2009年，蒙莫梅(Mon Mome)在英国国家赛马障碍大赛(the Grand National)中以100比1的赔率获胜之后，博彩公司都很高兴。**绝对风险，还是相对风险？**

假设在一群特定的人中，未来50年内大肠癌的发病概率是1/1000。我们预期在10 000人中，会有10个人罹患大肠癌。一种新的药物会将这个概率减小至1/2000：意味着如果药物投入使用，10 000个人中只有5个人会因癌症身亡。医药公司就会将新闻稿加上“患癌风

险降低50%”的大标题。这的确是精确的：对于每个人来说，风险的确会减半。

这个计算方法描述了**相对风险的降低**，并且经常会因为对数据的过于乐观的解读而饱受诟病。假设初始的风险是1/10 000 000：将它减小50%得到新的风险数字1/20 000 000，但是在前后两种情况中，这个风险是相当的小，以至于在10 000个人中，我们预期两种情况给出相同的患者数字——0。尽管风险被减半了，药品也几乎不会产生什么作用。

但是假设这群人患癌的风险是比较高的，比如说40%。一种能将概率减少到20%的药物不但有资格使用那个大标题，甚至会被称为重大突破，因为在10 000个人中，患癌人数整整减少了2000人。

考察绝对风险而不是相对风险的变化，通常是更有意义的。在上面的第一个情景中，绝对风险从0.1%变成了0.05%，降低了0.05%；在第二个情景中，降低了一个极小的数字0.000 005%；而在最后一个情景中，降低了令人印象深刻的20%。

一个合理的描述方式是陈述为了阻止一个恶性病例的发生，平均有多少个病人应该使用药物——**需治疗人数**(the Number Needed to Treat)或者说NNT。这个数字越小越好，它也就是绝对风险的变化的倒数。在上面的例子中，NNT分别是2000、20 000 000和5。

为了阻止一个恶性病例的发生而去治疗20 000 000人，很难说这是正当的。NNT，连同对治疗费用和疾病影响的严重程度的了解，让我们能做出关于分配卫生保健资源的合理决策。

组合小概率

在大量均具有微小概率的事件中，有多大可能至少有一个发生？在考虑灾难发生概率的时候，这会是一个相关的问题。如果无数组件中的任何一个失效，复杂的系统或者机器就会失效；两架飞机会相撞吗？一座核电站的堆芯可能会熔毁吗？所谓的**波莱尔-坎泰利引理**[2](Borel-Cantelli Lemmas)给出了一些提示。数学结果告诉我们，在很多情况下，重要的是这些微小的概率的**和**：如果这个和是无穷大的，那么灾难一定会发生。

结果就是我们永远不会满意于当前的安全标准。持续进行改进是必不可少的。因为无论我们的安全标准有多高，在某一个月中，一定会有不为0的失效概率：而且无论这个值多么小，如果它保持不变(或者它降低得比较慢)，大量月份的概率和就会变得无穷大，灾难一定会在某一时刻发生。

一个持续改进的充足计划也不会保证一定能避免灾难，但是满足于现状就是坐以待毙。

一些误解

(a)当一个医生告诉患者一种特定的药物有30%的概率会引起令人不快的副作用，他的意思是他预期有30%服用了这种药的患者会遇到副作用。然而，这个患者也许会相信当她服用这种药物的时候，在30%的次数中有副作用产生。医生考虑的是他看到的所有患者，这个患者考虑的是她服药的次数——他们的**参照类别**(reference class)是不一样的。

(b)公众会如何解读电视天气播报员说的“明天芝加哥下雨的概率是30%”？播报员们预期他们的听众会进行频率解读，即长远地看，如果天气情况类似于现在看到的那样，在30%的情况下，第二天会下雨。

但是当被问及此事的时候，即使在那些对于“30%概率”这个说法感到满意的电视观众之中，解读也是各种各样的。一些人认为他们被告知城市中的降雨区域超过30%；其他人认为芝加哥会在一天中30%时间中下雨；还有一些认为30%的气象学家预期会下雨! 一小部分人认为一定会下雨，30%这个数字意味着降雨的强度。播报员们提及的事件和听众们理解的事件会有很多不匹配。

(c)如果随机选择至少23个人聚集在一起，其中有两个人生日相同的概率就会超过1/2。人们第一次听说这个事实的时候通常会感到惊讶，但是在计算证明了这个断言是正确的之后，他们会被说服。然而，一小部分人仍然不会被说服，因为他们以为自己被告知，如果另外22个人被聚集起来，其中的一个和**他们**生日相同的概率超过1/2。听仔细!

(d)假设有一枚被认定是公正的硬币，连续9次投掷中均反面朝上。或许是受到了某些要求正面立即掺和到连续出现的反面之后的“平均法则”的影响，一些人就会宣称下一次投掷几乎确定地是正面朝上。这样的法则并不存在。大数定律的确会暗示正面和反面朝上的比例是相等的，但在**长远的**情况下，任何的连续9个反面的序列都会被前面或者后面的1000次投掷抵消。

或者一些人会(正确地)指出出现连续10次反面朝上的概率小于1/1000，所以当他们看到9次连续的反面之后，可能就会“推断”下一次也是反面是非常不可能的。但是这是伪逻辑。即使我们有连续的9次反面朝上，第10次反面朝上的概率也是1/2。这是事件的绝对概率和条件概率之间产生的混乱，它也在1996年的一个著名事件中体现出来：赛马骑师弗兰基·戴图理(Frankie Dettori)在雅士谷(Ascot)赢得了前6场赛马，因为从来没有人会赢得一天之中的7场赛马，戴图理赢得最后一场几乎“必须”是不可能的。但是他的确赢了。极少有人会在7

场比赛中的前6场均胜利，但是一旦有人做到了这一点，他可能也会赢得最后一场。

问你自己：我是在评估20件事情发生的绝对概率，还是已知前19件事已经发生的情况下，评估第20件的条件概率？

(e)报纸经常会在巨大的时间压力下出版发行，所以一些文章包含有荒谬的内容也容易理解。但是有3个报道是需要批评的。

当从超市买来的6个鸡蛋全都是双黄蛋，报纸就会宣称一件令人震惊的事件发生了。1000个鸡蛋中，只有1个有这样的特性，所以得到一整盒的双黄蛋的概率就是连乘6次后的一个非常小的值了。这个数字结果是如此小，以至于如果你每秒打开一个盒子，你预期要用300亿年才会遇到一个只装有双黄蛋的盒子。

但是这个计算只有在6个鸡蛋都是从双黄蛋比例是1/1000的大量鸡蛋中被独立选择而来的时候才有意义。但是事实并不是这样。在装入盒中之前，鸡蛋都是被按照尺寸分类了的。有些盒子上甚至被标上了装有双黄蛋的标签……

针对某一个英格兰足球队长私人生活的指控曝光，一名记者对于球队经理的4种可能的反应给出了“可能性”数值：

(a)将他从球队中开除——1/10；

(b)留在球队中，但是要求他辞职——3/10；

(c)留在球队中，但是废除队长职务——6/10；

(d)不采取行动——8/10。

这4个估计的概率中的任何一个单独都是貌似可信的。但是因为它们都关联着互斥的结果，它们的和就不能超过单位1。然而这些“概率”加在一起是1.8。

第3个报道中说，在英国国家彩票中赢得至少50 000英镑的人的名字，大多包括了约翰(John)、大卫(David)、迈克尔(Michael)、玛格丽特(Margaret)、苏珊(Susan)和帕特里夏(Patricia)。目前为止看起来还好，但是因此应该在购彩票集团中引入有这些名字的人就是很荒谬的了!

描述未知

从一个洗好了的牌堆中抽出一张牌，我预期红牌和黑牌是等可能的，我可以很自信地将红牌的概率赋值为“1/2”。使用一个稍微令人费解的方式，我会说我将抽到红牌的概率的**分布**百分之百地分配给1/2——百分百显示了我的信心。如果我确定概率是一个确定的数字，那么我就将这个数字附加上百分之百。

但是我经常无法选择单独一个数字：我对错过中转火车的最佳估计可能是3/4，但是2/3和4/5也貌似是可信的，甚至接近于极限值0和单位1的值也不是完全不可能的。我可以用一个0到单位1的连续分布来描述我对这个不确定概率的感觉。

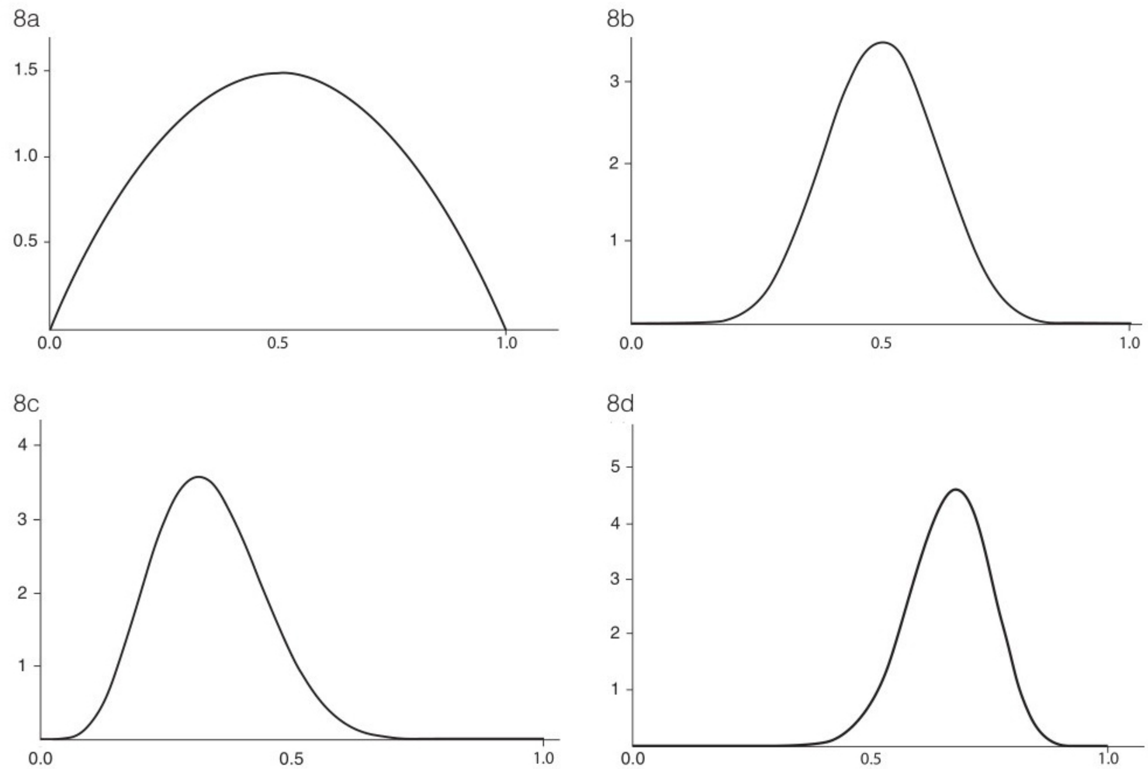


图8　贝塔分布的一组图像

完全未知的概率——非常罕见的情况——可以被图5中的连续均匀分布描述。更常见的情况是，首先有一个中间值，它是你以单个数字对这个概率给出的最佳猜测，然后你会直觉地认为更大和更小的值的分布会降到0。图8展示了具有这种性质的被称为**贝塔分布**(beta distribution)的一组图像。

图8a是我们对概率的值知之甚少的情况，我们将1/2当作最可信的值，但是像1/5一样小和4/5一样大的值也是很有可能的；在图8b中，我们对这个值接近于1/2更有信心了，同时也不排除极端值；在图8c中，我们最相信的值接近于1/3；图8d中的分布强烈地集中在了2/3周围，而且对于这个值小于1/2的期望是非常低的。

我的车加了10升油的情况下能开多远？当预热之后匀速行驶的情况下，我预期能开90英里，但如果我在几周内进行过几次短程行驶，就更可能是60英里了。在这两种情况中，这个距离都会有一些不确定性，可以用连续分布来表示。为了找出是哪种分布，想象一下将这些汽油分成10毫升一杯的小份。一共就会有1000个这种杯，总的距离就会是这1000杯中每一杯行驶的距离之和。回忆中心极限定理，大量随机量的**和**趋向于遵循高斯分布。

对于恒定速度的高速公路行驶，我会选择中心位于90的高斯分布，方差小所以峰很尖。对于在城镇中的情况不定的行驶，我也会用高斯分布描述，但是中心在60而且分布比较宽来显示较大的不确定性。

效用

有位仙子让你进行一次选择。她给你1英镑，或者她掷一个硬币，如果你猜对了正反面，她就会给你10英镑，否则你什么都得不到。你会做出什么选择？

备选的选项是确定得到1英镑或者一个公平的赌局，回报是5英镑。几乎所有人都会更喜欢后者，但是将问题中的金额都乘上1 000 000，人们的选择就必然会变化。比起什么都得不到与得到500万英镑概率对半来说，确定能够得到100万英镑，更有吸引力。这种差别背后是**效用**(utility)这个概念起作用。

对较小的总金额，翻倍**的确**意味着翻倍，但是如果100万英镑会为你和你的家人带来相当程度的快乐，将这个值翻倍就不会使快乐翻倍。无论你选用什么单位来计量货币金额的“价值”，通常效用和金额之间的关系曲线的形状会像图9所示：图线总会上升，一开始像一条直线，但后来增长得越来越缓慢了。

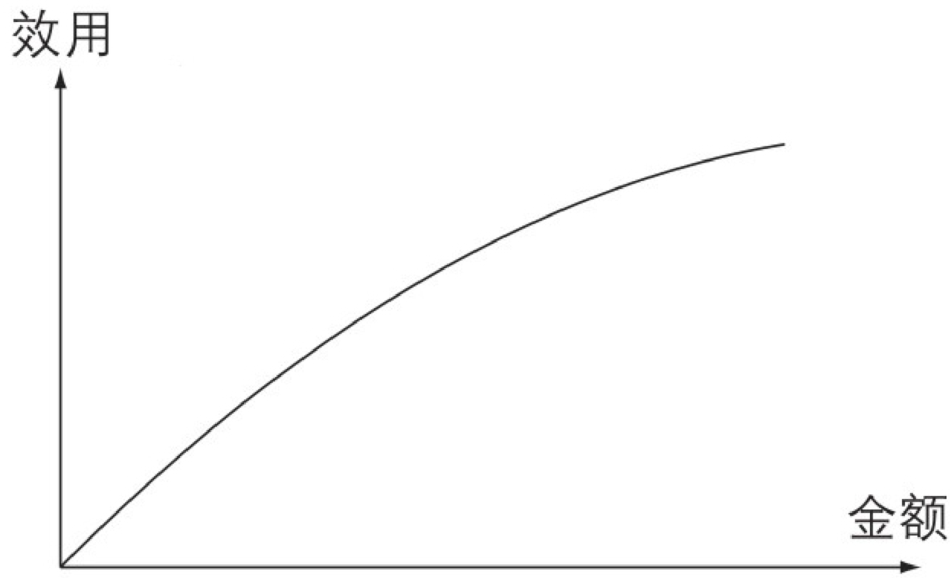


图9　一般的效用曲线的形状

“效用”解释了为什么房主和保险公司会同意总额250英镑是合理的年度保险费，这笔保费能够确保你在火灾、沉降或者洪水中免受25万英镑的房屋损失。保险公司的大额赔款意味着，对于任何单独一栋房子，这笔赔款可以看作与相关的数目较小的赔款预期的和是相等的。只要在任意一年中需要赔款的概率小于1/1000，保险公司在这笔交易中的期望值是正的，它就会盈利。另一方面，没上保险的房主就需要面对巨额的负效用，如果房子遇险，损失25万英镑，所以对于她来说，主动放弃250英镑来阻止这种可能性就是一笔合适的交易。

给诸如电视机、微波炉等的损坏上保险通常是坏主意。损失的总额很小，负效用和赔款本质上是相等的，保险公司会收取大额的保险费来保证盈利。应该将假想的保费存入银行作为修缮基金，而不是买高保费保险。很少有人会因为这么做而后悔。

如果你能成功地创建你自己的效用函数，你能够用它在不确定性存在的情况下做出选择。对于每一项行动，计算各种结果的**预期效用**

(expected utility)，即用对应的概率对效用加权，然后选择预期效用最大的一项行动。

无论具体情况如何，这就是概率论者做选择时候的通用秘诀。

[1] 全称Epsom Derby，为每年在英国举行的赛马比赛。

[2] 概率论中的一个基本结论。

06 人们玩的游戏 Games People Play

许多休闲游戏(recreational games)都是能力和机会的组合。能力是能提升的，但机会就与运气有关了。对于这里讨论的所有“游戏”，你会很容易相信它们是有限多个结果，并且它们都是等可能的。因此在这一章中，除非明确声明，我们都会使用处理概率的经典方法：计算可能结果的个数，任何事件的概率就是它们发生的结果占总结果数的比例。

我的目标是展示在不确定性存在的情况下，概率能如何帮助一个玩家做出正确的选择。观众对概率的理解也会增加观赏游戏的享受和乐趣。

彩票

在英国国家彩票中，最常见的模式是6/49。涂有不同号码的49个橡胶圆球在塑料桶中被搅拌打散，然后随机选择出6个来。赌博者花费1英镑来选择6个号码，如果这个选择中包含了至少3个中奖号码，他们就会赢得奖金。但是因为只有50%的彩票收入会成为奖金，彩票玩家平均收入的值远低于赌场赌博或者赛马。

彩票对人们的主要吸引力是巨额的中奖预期，虽然它们很难得到——曾有一张英国彩票开出2200万英镑的大奖，美国彩票的奖额已经超过了3亿美元。计算可知，用一张彩票中得头奖奖额的概率，在英国大约是1/14 000 000，在欧洲百万彩(the Euromillions Game)中小于1/116 000 000，在美国超级百万(the USA Mega Millions)中大约是1/176 000 000。

为了体现这些概率到底有多么小，我们单看英国彩票。实际上，一个随机选出的40岁男子在一年内死亡的概率是1/1000。所以他在一天中死亡的概率是大约1/365 000，在一小时内死亡的概率是1/9 000 000，所以概率低到1/14 000 000表示的就是他在接下来的35分钟内死亡的概率了! 对于美国超级百万彩票来说，在相同的假设下，这名男子中得头奖的概率和他在接下来3分钟内死亡的概率几乎相等。

尽管有极低的回报和几乎没有的赔率，“效用”能够给出买彩票行为的合理解释。用1英镑作为交换，无论如何你都会得到平均50便士，另外50便士你就用来买了梦想未来奢侈生活的权利、做了慈善事业，还有像我一样确信你在浪费钱的人的潜在的羡慕。这些权利的确算是一些效用。

我们应该假设未来的彩票开奖结果和先前的结果是相互独立的——一个无生命的橡胶球不会“记住”它是否“应该”被抽中。除了作弊，没有任何方法能够改变中奖的概率。但是你**的确可以**影响你可能获得的奖额的大小，因为在这些彩票规则中，奖池确定比例的部分会在每个奖励等级中被相应的中奖者们平分。我们有了一个训练某种能力的机会。

某些很小或特别的确定的数字(例如生日)会比其他数字更经常被选择，并且许多彩票玩家将自己的选择平均分配到自己买的彩票中，可能是误认为这么做就进行了“随机”选择。结果就是，几个较大数字，或者相距不远的一串 数字，或者边缘的数字，比较不经常被选择。如果你能够确定其他的玩家做出了什么类型的选择，并且做一些**不同的**选择，你中奖的概率不会受影响，但是一旦你中奖了，你就会赢得比平均值更多的奖金。

别基于“没有其他人会想到这个”耍小聪明，例如选择{1， 2， 3， 4， 5， 6}，或者选上次开奖时候的中奖号码。他们的确会想

到。在英国国家彩票首次开奖的时候，大约10 000人选择了前6个数字。在2009年9月，相邻两次保加利亚彩票的中奖号码完全相同：第一次没人选那个号码，但第二次就有18个人选了。

已知其他玩家会像他们以前做的那样继续频繁地选择他们特定的号码，对于英国的6/49类型的彩票来说，下面的过程会帮到你。取普通的52张一副的纸牌然后去除3张。将剩下的纸牌与数字1到49一一对应，洗好牌，然后抽取6张纸牌。这是一种完全随机的抽取6张纸牌的方法。人类不可能不借助外物进行这种随机选择，我们需要这类辅助。

然后对于这6张牌，作如下规定：

(a)它们加在一起至少达到177(为了偏向较大数字)；

(b)在写在彩票上的时候，它们分成了2、3、4或者5个集群；

(c)它们中有3、4或者5张的值落在彩票数值的边缘上；

(d)它们不会在彩票上形成任何明显的规律。

如果其中任意一个条件不被满足，就将这6张纸牌放回牌堆，彻底地洗一遍牌，然后重复这个过程。

即使你遵循这个秘诀，你仍然预期会输钱——因为全部的回报只占彩票公司收款的50%这个事实是没法克服的。但是这样你就不太可能同整个世界分享你的头奖了。

电视游戏

“黄金球”(Golden Ball)在2007年首播。最后两位玩家会面对11个金球，其中一些对应一些奖额，另一些叫杀手球，不对应任何奖额。玩家从中选取5个球，生成潜在奖金，选中的任何杀手球会将之前

选取的球代表的金额减小到原来的1/10。因此在选取了一个50 000英镑的球之后再取两个杀手球就会让奖额变成500英镑。

所有的球表面上看起来都是一样的，所以玩家完全是随机选择。从11个物体中选取5个一共有462种方式，所以从中选取到5个最有价值的球的概率就是1/462。在最初的288期节目中，这只发生过1次。

假设一个球名义上价值1000英镑：即使不考虑杀手球，选中它的概率也只有5/11，所以它的真实的价值就是455英镑。选取任意杀手球都会降低总奖额——在有3个杀手球的情况下，平均值的计算结果是255英镑。

当选取完5个球，实际的奖金已经知道的情况下，两个玩家都会做出一个彼此保密的决定，或是将自己的奖金**分享**给另一个玩家，或是**夺取**所有的奖金。他们俩会同时公布他们的选择：如果都选了分享，那么两个人分享奖金；如果其中只有一个选择了夺取，这名玩家就会获得所有奖金；如果两个人都选了夺取，两个人都不会得到任何奖金。

这是在博弈论中非常著名的一个情景，它被称为“**囚徒困境**”(the Prisoner’s Dilemma)。假设你的对手选择了分享，那么你选择夺取就会获得较多的钱。如果你的对手选择了夺取，你怎么选都无所得。所以无论另一个人选择什么，你可以说选择夺取就永远不会输。经常发生的情况是，两个人都选了夺取，那么唯一的赢家就是不出任何钱的电视节目制作公司。

不同版本的“成交与否”(Deal or No Deal)已经在70多个国家播放。在英国，22个密封的盒子装有从1便士到250 000英镑不等的钱。这些盒子被随机分配给22个玩家，其中一位叫艾米的玩家将会在这一天参与游戏。她自己的盒子直到游戏结束都一直是关着的。她首先选

择5个其他的盒子，之后盒子中的金额会被展示出来。庄家这时会出价来交换艾米的盒子中的金额。要是接受的话，她说“成交(Deal)”，则游戏结束；说“不成交(No Deal)”，则拒绝这个出价。若游戏持续进行，更多的盒子会被打开，庄家会开出新的出价，等等。

出价时，仍在游戏中的确切金额是已知的，所以他们的平均值是很容易计算的。在最初的几轮中，正常情况下庄家的出价会远远小于这个均值，但是艾米必须将她自己的效用函数牢记心上：如果她十分想要5000英镑，而出价是4500英镑，她就应该理智地接受，即使剩余盒子的平均金额超过20 000英镑——如果坚持不成交的话，她可能最后只得到1便士。

22次中有1次，艾米会被分到最高金额的盒子，但是她很少会赢得那么大的金额。效用可以给出一个令人信服的解释。在最后一次决定中，剩下两个盒子，其中一个是250 000英镑，另一个也许是2英镑。如果庄家出价80 000英镑，虽然这比均值125 001英镑小很多，也只有最有勇气的或者最富有的艾米才会拒绝。千鸟在林，不如一鸟在手……

已知庄家总是会给比剩余的盒子的平均金额少的出价，长远地看大数定律保证了“成交”的选手获得的金额比他们盒子中的少。所以庄家的确会支付出价，但是也会在选手成交之后获得盒子中的金额，就会长远地获利。

“金钱本色(The Color of Money)”被认为是最令人紧张的电视节目，它在2009年只播出了4期。但是它给出了计算概率过程中应用加法和乘法定理的极好的例子。

1000英镑、2000英镑，一直到20 000英镑被随机分配到20个不同颜色的盒子中，玩家葆拉要达到一个预先设定好的目标金额，例如64

000英镑。为了达到这一点，她可以选择至多10个盒子，一次选一个。如果她(不知情地)选择了14 000英镑的盒子，1000英镑、2000英镑，一直到14 000英镑的数字就会以稳定的速度在屏幕上变换，每显示一个数字她都可以叫停。如果她及时叫停了，就会积累下最后显示出来的那个数字，但如果她等得太久，就什么都攒不下。如果在选取了10个盒子之后，她还是没有达到既定目标，她就什么都赢不到。她应该采用什么策略呢？

除了颜色，所有的盒子都是相同的，所以葆拉从每一轮中剩下的盒子中做选择是完全随机的。在她的最后一轮中，剩下了11个盒子，她的策略显而易见：例如，如果她想要9000英镑来达到目标，而且恰好6个盒子价值9000英镑或者以上，她就会计划在9000英镑出现时叫停，她成功的概率是6/11。但是在早期的几轮中她应该怎么办呢？

或在还有两轮的时候，剩下的12个盒子装有(以1000英镑为单位)1、4、5、6、9、10、12、13、15、17、19、20，而她需要额外的15 000英镑。当她看到7000英镑时叫停是不合理的；如果7000英镑曾经出现过，她就知道这时候盒子中至少有9000英镑，所以她就应该在9000英镑时叫停，显然这是更好的策略。她也可以将自己的选择限制在这剩下的12个数字中。相同的论证也同样适用于早前的几轮中——她叫停的最佳时机总是在剩下的某一个盒子里的金额出现的时候。

如果葆拉的确想要在9000英镑出现的时候叫停，她可以这么说：“12个盒子中有8个至少有这么多，所以我成功的概率就是8/12。而且如果我这轮的确成功了，在最后一轮中我只需要6000英镑，这时候11个盒子中有8个是我想要的。乘法定理告诉我们这两件事情同时发生的概率是(8/12) × (8/11) = 64/132。第一轮中有4个盒子的金额少于9000英镑，所以第一轮我什么都得不到的概率是4/12；这时我在最后一轮中就需要15 000英镑了，这个概率是4/11。再一次根据乘法定

理，这种情况的概率是(4/12) × (4/11) = 16/132。这两个胜利是互斥的，所以加法定理告诉我们胜利的总概率是80/132。”

她也可以对她其他的选择策略进行相似的分析，例如在第一轮中期待6000英镑或是12 000英镑。我请你来做这些计算——附录中描述了她的最佳选择。

在这个电视节目的筹划阶段，有人出过一个主意，请一位专业数学家来提供及时的策略建议。葆拉可能会讲她会在8000英镑的时候叫停，这名专家也许会说：“这个选择不错。如果你这么做了，你赢钱的概率就是75%。但是如果你计划在11 000英镑的时候叫停，你的成功概率就会是80%。”

你可以想到会发生什么! 专家说得都是对的，但如果葆拉更改了策略，却最终失败了，同时的确她最初的直觉会奏效。就一定会有一些小报叫嚷：“数学研究员夺走了军事英雄遗孀的64 000英镑。”

在我们调查过电视游戏节目中的数学之后，得知他们从未引入过数学角度建议员，我们都松了一口气!

纸牌游戏

大数定律意味着我们会在长时间的游戏中公平地得到好牌或者坏牌，所以最终游戏结果展现的是能力水平。我们来看3种流行的纸牌游戏。

在“二十一点”中，庄家必须遵循关于何时发牌的固定规则，玩家想做什么就做什么。直到爱德华·索普(Edward Thorp)开始赢得数额巨大的赌金之前，赌场运营者们都相信没有任何其他系统能够击败他们内在的优势。他们的逻辑中存在一个致命的瑕疵：虽然他们可以期望在牌堆中有6或者8副牌的时候赢得奖金的1%~2%，但是几局过后局

面可能会转向对玩家有利。赌场忽略了使用剩下的牌的**条件概率**，而只考虑到了由完整的牌堆计算得到的概率。

索普开发出了一个追踪牌堆中剩余牌的方法。当数值较大的牌占比较高的时候，可能规则会变成强迫庄家抽牌，最终导致其点数超过21点而爆牌输掉。在相同的情况下，玩家却可以选择不抽牌。只要牌堆中数值较大的牌的比例较低或者适中，索普就会下尽可能小的赌注，而当牌堆组成对他更有利时，他会下更大的注。简单，但是高效。

当牌堆的组成的确对玩家有利的时候，他应该下多少注呢？约翰·凯利(John Kelly)在索普分析研究的几年前给出了精确的答案：下注的资金应该与他的优势大小一样。这个策略会让他资金增长的速率最大。

例如，假设他有1000英镑而且游戏对他有一点点有利——他获胜的概率是51%，失败的概率是49%。他的优势就是2%，所以他下注的比例应该是总资金的2%，即下注20英镑。下一次，他就会有980英镑或者1020英镑，如果他的优势保持在2%，按照出现的不同结果，他应该下注19.60英镑或者20.40英镑。如果他过于贪婪——在凯利指出比例仅2%时下注总资金的10%，尽管他有一定的优势，他最终还是会破产。他的资金是有限的，下的赌注过高以至于不能承受不可避免的失败的趋势。

赌场会采取措施来识别和禁止熟练的数牌者，从来没有人因理解概率的出色能力而获利。

我们注意到，在考察法庭上的证据会如何改变我们相信有罪或无罪的程度时，正确的方法是采用贝叶斯公式。在例如惠斯特桥牌的纸牌游戏中，使用这个规则会提升你做出正确决定的概率。简便起见，

我使用法律相关的词汇，用“有罪”来表示一个特定的对手同时持有一组特定的牌，比如说红桃Q和K，同时用“无罪”表示她至多持有这些牌中的一张。

使用计数的方法，我们可以找到她持有全部两张牌的情况占全部情况的比例，这样就会给出对于“有罪”概率的最初评估。按照标准的习惯，最好将这个概率转化成它等效的赔率。最开始的计算完成后，我们可以说我们得到了(有罪的)**先验赔率**(Prior odds)。

当纸牌游戏进行下去，相关的**证据**就会浮现出来——但她或许是在耍花招。为了考察这些证据是如何影响“有罪”的赔率的，我们来计算一个叫作**似然比**(Likelihood Ratio)的量。首先，假定“有罪”(她同时持有K和Q)，评估这些证据出现的概率；然后，假定“无罪”(她至多只有其中一张)，评估这些证据出现的概率。似然比就是第一个值和第二个值的比值。

我们现在就可以推断出**后验赔率**(Posterior odds)，也就是说在考虑到证据的情况下“有罪”的概率，使用贝叶斯公式，也就是：

后验赔率=先验赔率× 似然比

这个定律的形式是清晰直白的：似然比越大(也就是说，你的对手的确是有罪的时候，证据越容易出现)，有罪的赔率就越大——但是这个公式精确地告诉你相关证据如何影响有罪的概率。

要看这个理论是如何运行的，来思考一个现实的情况：我们的对手要么是同时持有K和Q(有罪)，或者她只持有K(无罪)；先验概率告诉我们这些选择差不多是同等可能的。如果她是有罪的你最好就出A牌；如果她是无罪的，你就应该出一些其他的牌。证据这时出现了，她出了K。

如果没有证据，你就只能凭空猜测，而且你也就有一半的概率会赢。如果她是无罪的(她只持有K)，那么证据(她出K)出现的概率就是100%；但是如果她是有罪的(她同时持有K和Q)，她可能也就会出Q而不是你看到的K，所以证据出现的概率就是50%。这两个数字比值是1/2，所以贝叶斯公式告诉你后验概率就是1/2。也就是说，她无罪(只有K)的可能性是有罪的2倍。所以现在出A就有2/3的概率获胜。

在正确使用概率的情况下，你应该希望能做得更好，如果你能够以2/3的概率获胜，就别以1/2的概率。你不能保证每局都赢，但是这样做你可以提升获胜的概率。

桥牌玩家把这种情况称为“限制选择原则(the Principle of Restricted Choice)”——如果对手只有一张K她就必须打出，如果同时有K和Q，那她可以做选择。她的确打出了K的这个事实，让概率向她不得不这么偏移了。

如今，最流行的扑克的形式是德州扑克[1](Texas Hold Em)。每个玩家发2张牌并且设法使自己手里面的牌被最有效地打出，5张公共牌随后会被陆续发放。公共牌发放之后，为了最有可能击败其余两组牌，哪一组牌是最好的？

手牌A：**两张**梅花2，黑桃2；

手牌B：黑桃A，方片K；

手牌C：红桃J和10。

这当然是一个陷阱问题：在仔细计算之后，手牌A会以52%的概率击败手牌B，手牌B以59%的概率击败手牌C，手牌C以约53%的概率击败手牌A。所以你就会选择A而不是B，选择B而不是C，但是你同时也会优先选择C而不是A! 如果你的对手从这3组牌中选取任何一组，随后你在剩下的牌组中随机选择1组，你都有超过50%的概率获胜。

扑克游戏需要的远远不止概率技巧。你必须对对手有可能持有的牌，以及你什么时候应该虚张声势做出评估。但是有些时候概率十分有用。假设赌注总额是50个筹码，还有1张公共牌等待发放。这时你发现，如果最后一张牌是黑桃，你就能打出同花顺，这样你一定会赢；如果它不是黑桃，你的一个对手就会赢。你应该在游戏中押下更多的筹码吗？

忽略掉你已经在赌注总额中放了多少筹码。重要的是未来的情况。你能看到6张牌的牌面——2张在你手里，4张在桌上的公共牌中。在46张未知的牌中，9张黑桃会让你赢得胜利，其余的会导致失败。在赌注总额已经是50个筹码的情况下，再多付10个筹码来看最终分发的纸牌是值得的吗？再付20个呢？

通过计算得到你一定要多押的x个筹码带来的平均收益(或者损失)，你可以得到一个阈值x。它将会从长远角度带来收益。附录中给出了这个问题的答案。

[1] 简称德扑，最流行的扑克衍生游戏之一。

07 在科学、医学和运筹学中的应用 Applications in Science, Medici ne, and Operations Research

我们会根据情景使用不同方式来对概率进行评定或者诠释。但是，就像大卫·汉德(David Hand)在他的《 牛津通识课：统计》(Statistics: A Very Short Introduction)中写的那样，“……微积分是一样的”，换言之，概率的操纵方式是不变的。

你头脑中要牢记这个学科的中心思想：加法和乘法定理、独立性、将客观概率和频率联系起来的大数定律、在将随机数求和时候使用的高斯分布、其他的一些经常出现的分布函数、反映总体情况时有用的平均值和方差。我们可能不指望我们对相关概率知道得像前几章那样精确，但是一个对于问题大致正确的回答对于作出合适的决定有良好的指导意义。就像统计学家乔治·博克斯(George Box)所说的那样，“所有的模型都不是完全正确的，但是有一些是很有用的”。

下两章中举例说明了概率的应用，这些应用以章节标题粗略地分了组。

布朗运动和随机游走

1827年，植物学家罗伯特·布朗(Robert Brown)观察到，在液体中悬浮的花粉粒子似乎随机地在动来动去。将近80年之后，阿尔伯特·爱因斯坦(Albert Einstein)对其给出了一个解释：花粉粒子被液体中的分子持续地撞击。这种运动当然是发生在三维空间中的，但是为了创建一个令人满意的模型，我们首先思考在一条直线上的运动。

假设每一步运动都是具有固定长度的，有时向左有时向右，每一次运动都是独立的。这个概念就叫作**随机游走**(Random Walk)。在许多次跳跃之后的位置只取决于向两个方向的跳跃次数的差值；从起始点计算的距离平均值和方差与进行跳跃的次数成正比。

下面进行一个微妙的计算：在一个固定的时间段中，**增加**跳跃的频率，并且**降低**每一次跳跃的距离。在这两种因素平衡的情况下，极限情况就是连续运动，运动经过的随机距离遵循高斯分布(依据中心极限定理)，这个分布的平均值和方差都与时间段长度成正比。如果向左和向右的运动是等可能的，平均值就会是0。

爱因斯坦对于布朗的观察作出的解释，是粒子在三维空间中运动，基于上文给出的原因，在每一个方向上的运动都遵循高斯分布。他对原子和分子的行为作出了预测，这些预测推动了一些实验，这些实验消除了有关原子和分子存在的所有悬而未决的问题。

术语“**布朗运动**(Brownian motion)”本应该专指在液体中的粒子的实际运动，但是它也被用于指代描述这种运动的数学模型。

随机数

“随机数”这个词指代的是下列两种想法之一。第一，就像理想情况下的色子游戏或者轮盘赌游戏，一个数从一个有限的列表中被取出来，所有的这些数都是等可能的。第二，就像从一个随机的一点折断一根木棍一样，一个点被从一个连续的区间中取出来，没有任何一小段是比其他的小段更容易被取点的。取出这样的随机数中的一段长序列的方法具有广泛的应用，序列中的每一个值与其他的值都是相互独立的，下一部分会举例说明这一点。

在1955年，一本名为《 一百万个随机数》(One Million Random Digits)出版了。它的确就是它的书名描述的那样：一页又一页的0到9的数字，被分块以便阅读，但是连续的数字是完全不可预测的——无论前面的数字是什么，你都会有1/10的概率猜对下一个。如今，现代计算机已经有内置的软件来获取与之相同的结果。输入一个初始值——随机数**种子**(seed)，一个确定的数学公式就会给出下一个值，它被当作新的随机数种子，以此类推。这个过程中毫无随机意义可言，并且如果每一次都是用相同的随机数种子，就保证会生成完全相同的数列。但是，基于数学公式的巧妙选择，生成的序列成了统计检验中的基础，而且对于任何目的和用途来说，它**看起来**就好像是随机的一样。我们用**伪随机数列**(pseudo-random sequence)这个术语来称呼它。

无论在这个过程中有多么小心，一定会有一些方法中的谬误将会引起对随机数应用场景的担心。但是依赖大量受人尊敬的科学家的经验，我还是会认为我电脑中根据需求产生的随机数是可以接受的。(显而易见的内部人员欺诈的风险导致了这些方法不能应用于彩票摇奖，或者英国溢价债券。)

蒙特卡罗方法

连续37次转动标准的欧洲轮盘赌轮会得到多少不同的数字？理论上讲，这个个数会是1~37之间的任何数，但是那些极端的值会十分罕见，最常出现的不同数字的个数是多少？

这个问题第一次呈现在我面前的时候，我没有立即看出有什么简单的方法能解决它。转动赌轮37次会得到37^{37}(一个有59位的十进制数)种可能的结果，而且当你试图罗列，例如有28个不同数字的结果的时候，你将会很快失去热情。一个更加吸引人的方法是进行一个所谓的**蒙特卡罗模拟**(Monte Carlo simulation)。

在这里，计算机产生的随机数流将会用来模拟37次转动赌轮的结果，之后计算机将会计算有多少不同的数字出现了。这个过程将会重复100万次，结果是24个不同数字出现了203 739次，而23个数字只出现了199 262次。最接近的竞争对手是22或者25个数字，都出现了不到160 000次。大数定律告诉我们不同的结果出现的频率将会稳定在它们的各自的概率，而且这些数字的确本质上证实了这件事：最有可能的结果就是有24个不同的结果会出现，概率刚刚超过20%。

几天后，我羞愧于当时没能找到一种标准的方法来解决这个问题!我能够计算出对于任意的X，转37次赌轮得到X个不同数字的精确的概率，以证明上面的结论是正确的。但是这不会让应用于这类问题的模拟失去效力——快速而粗糙的结果也会是很有用的。的确，模拟给出的结果与精确计算得到的结果保持一致这个事实，增加了我对计算机随机数生成器按预想运行的信任。

一个更加严肃的蒙特卡罗方法的应用是在聚合物化学(polymer chemistry)中。一个分子是由大量的原子被随机扭曲的长链连接构成的。原子们只能出现在均匀分割的晶格中，并且关键的一点是，没有两个原子能够出现在同一个位置。从分子的一端到另一端的距离可能有多远?

我们可以认为原子被放置在一个醉汉走过的位置，这个醉汉摇摇晃晃地随机经过三维空间中的晶格，但是因为某些原因不能在同一个位置经过两次。没有不能重复经过同一个位置的这个要求，数学家们可以给出很好的解答，但是这个限制条件似乎将问题复杂化到了理论无法解决的地步。然而，哪怕一个不专业的计算机程序员也可以写出一个对这个复杂、曲折的链的合理的模拟，而且在100万、1000万，甚至10亿次重复之后，得到一个所需的精确答案。(回忆棣莫弗的工作，精确度与模拟次数的**算数平方根**成正比。)

假设你想要估计一片不规则形状叶片的面积。画一个包围了这片叶子的矩形框，然后模拟大量的随机分散到矩形内部的点的位置。你的估计就是用矩形区域的总面积乘以落在叶子边界内部的点的占比。

作为最后一个举例介绍的应用，假设保罗(Paul)想要建一个新的加油站。如果他安装4个加油泵，这是最小的可行的泵数，就会有至多8辆车的等待区；每个额外的泵减少2个等待区，所以如果他安装了最多的8个泵，就没有等待区了。为了计算多少个泵会让他的收益最大，他可以进行对安装了4、5、6、7或者8个泵的情况的模拟。

他应该知道潜在的顾客前来加油的比例和一辆车停在泵旁边加油的时间的分布，还有安装费用、运行费用与边际收益。他也应该考虑到如果没有加油泵空闲，或者队伍过长时，一个潜在的顾客直接开过去不来加油的概率。找到或者估计所有这些数字都是相对简单的，而且用计算机进行模拟会比在几个月中以不同的泵数进行实地试验**便宜很多**。因为他可以每一次都使用相同的随机数种子，他就可以在精确相同的情况下运行所有的模拟，对比不同的估计从而增加收益。

为什么叫“蒙特卡罗”呢？除了随机数和赌场游戏之间的明显的联系，这个名字其实原本用来指代军事领域对这些方法的应用，这其中就包括了早期的核弹的研制[1]。

电码中的错误

摩尔斯电码(Morse code)告诉我们如何只用两种符号，比如说0和1来传输消息。但是其中一些符号也许会被损坏，以至于原本发出的0接收到的时候就变成了1，反之亦然。甚至在较低的错误率下，接收到的消息也会与发送的消息有天壤之别。我们如何应对这种情况？

假设每个传送的符号都有相互独立的较小概率被损坏。我们可以重复发送这些符号，但是稍微一想就会发现，发送00和11而不是0和1根本不能解决问题：如果01或者10到达，确认到底传输的是00还是11就全靠猜测。我们会猜对一半，但是重复发送符号意味着我们可以预见其中会产生两倍的错误，所以两种因素大部分相互抵消了。但是我们考虑一下传输000和111而不是0或者1。

采用“少数服从多数”的原则来进行解码，所有的{000， 100，010， 001}都被理解为0，其他4种可能的情况被理解为1。如果只有1%的发送的符号被损坏了，那么当000被发送出去，利用二项分布告诉我们有99.97%的概率上述的4种序列被接收到。这意味着错误率从1%降到了0.03%，降低了超过30倍。如果每个数字重复5次，我们可以得到更好的结果，但是消息变长会增大开支。最佳选择将会依赖内在的错误率和传输的速度。

羊膜穿刺术

准父母(同时也是统计学家)樊娟娟(Juanjuan Fan)和理查德·莱文(Richard Levine)在考虑樊娟娟要不要接受羊膜穿刺术(amniocentesis)——可以检测她的胎儿是否患有唐氏综合征[2](Down’s Syndrome)。他们的经历可以作为其他类似情况的模板。

基于樊娟娟的年龄和简单的血液检查，我们可以给出胎儿患有唐氏综合征的概率为1/80；超声图像检查的结果是令人满意的，借助贝叶斯公式计算后，得到的患病概率减小到了1/120。羊膜穿刺术是一个侵入性的手术——一根中空的针刺入腹腔抽取羊水样本；如果作为唐氏综合征特征的21号染色体被检测出多余复制，就一定能够确诊，但是羊膜穿刺术这种检测有一定的风险导致流产，在这个情境中风险概

率估计是1/200。如果确诊胎儿为唐氏综合征后，父母们一定会选择流产。他们应该接受这个检测吗？

樊娟娟和莱文通过使期望效用最大化的逻辑分析过程得出他们的决定。可能出现的最坏结果是没有患唐氏综合征的胎儿流产，其效用赋值为0；最好的结果，是出生的胎儿不患有唐氏综合征，其效用赋值为单位1。不进行羊膜穿刺术，生下来患有唐氏综合征的孩子，赋值效用x；进行了检测，的确检测出了唐氏综合征的效用y应该大一些。(最后一个情况中检测导致的流产就无关紧要了，因为无论如何胎儿都会被流产。)

他们计算出了进行检测和不进行检测的期望效用。如果第一个值超过第二个，就应该进行检测，在这种情况下，就是要求y大于(119/200) + x，粗略地要求y大于0.6 + x。

如果樊娟娟和莱文认为确诊唐氏综合征之后进行流产的效用小于0.6，那么进行检测就会毫无意义。他们给生下尽管患有唐氏综合征的小孩的效用x赋值越高，那么y的阈值就会越高。如果这个效用x超过0.4，那么他们就一定不会接受检测。

选择x和y的合适的值需要一定的思考，而且如果基本的参数——接受检测的流产概率为1/200，不接受检测生下唐氏综合征婴儿的概率1/120——发生变换，最后计算得到的判断准则就会变化(参见附录)。简单地讲，如果胎儿患有唐氏综合征的概率小于意外流产的概率，接受检测就不合理了。真的是这样吗？

樊娟娟和莱文讨论了他们面临的难题，他们赞同的效用值的选择让他们决定接受检测。结果是美好的：没有多余的染色体，也没有发生流产。

血友病

血友病(haemophilia)是一系列伤口产生后无法凝血的疾病的统称。控制凝血的基因位于X染色体上，这个基因异常的概率小于1/5000。因为女性有两个X染色体，所以她们只会在两个染色体上的基因均异常的时候才会发病，发病概率小于1/25 000 000，但是男性只有1个X染色体(和一个Y染色体)，所以几乎所有的病例均为男性。

如果男性患有血友病，这在他做父亲之前就已经确定了，但是一个女性有可能有一个正常的X染色体和一个携带异常凝血基因的X染色体。这些女性被称为携带者，而且她把异常基因传递给孩子的概率是50%。女儿继承了异常基因会成为携带者，儿子继承了异常基因就会成为血友病患者。维多利亚女王(Queen Victoria)肯定是一个携带者，因为她的儿子利奥波德(Leopold)是一个血友病患者，而她的5个女儿中至少2个是携带者，她的另外3个儿子没有患这种病。

假设贝蒂(Betty)有一个患有血友病的兄弟，贝蒂有7个孩子，其中包括安妮(Anne)。那么安妮是携带者的概率是多少？

为了解答这个问题，得到贝蒂是携带者的概率就足够了，安妮的患病概率一定是这个值的一半。因为贝蒂的兄弟是血友病患者，所以贝蒂的妈妈是携带者。但就这个信息而言，贝蒂是携带者的概率是50%。而且如果安妮的**任意**一个兄弟患有这种疾病，贝蒂就一定是携带者，所以我们来看安妮的所有兄弟都正常的情况。

贝叶斯公式非常适合于这种情况。设“有罪”意味着贝蒂是携带者，因为有罪的先验概率是50%，先验赔率就是单位1。如果贝蒂是无罪的(不是携带者)，证据(所有的兄弟都是健康的)的概率就直接是100%。但是如果贝蒂是有罪的，每个安妮的兄弟都独立地有50%的概率避免错误的基因，所以每个健康的兄弟都会让似然比减半。将后验赔

率转换为后验概率，贝蒂是携带者的概率就相继是1/3、1/5、1/9、1/17……，对应于她有1、2、3、4……个全部健康的兄弟。

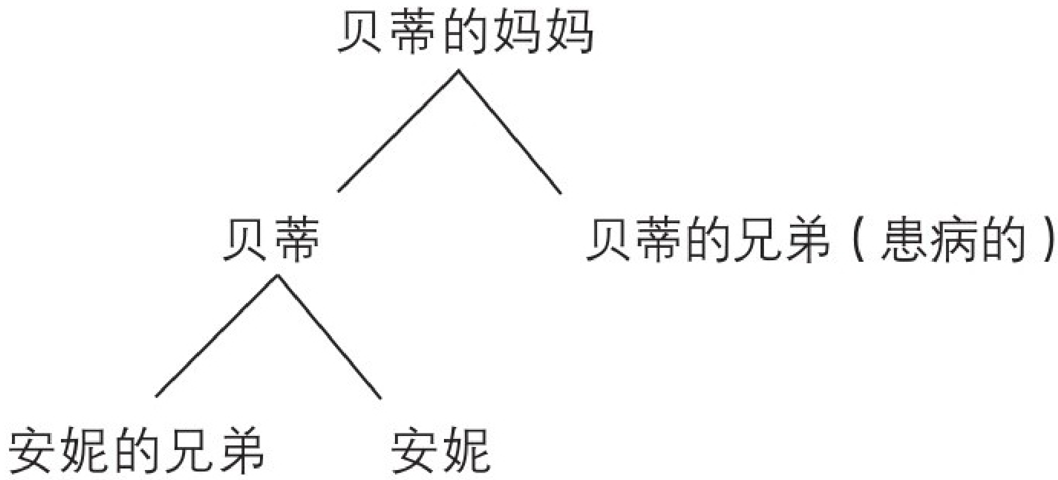


图10　家庭成员关系

作为消遣你可以思考一下，假设安妮同时有姐妹，她们其中一些有儿子，而且安妮的这些侄子们都不患病。这将如何影响安妮是携带者的概率？将你的答案与附录中的相对照。

流行病学

术语**群体免疫**(herd immunity)表示一个事实：如果足够多的人免疫了，那么即使传染病进入社群，流行病也不会发生。因此即使那些没有接种疫苗的人也非常不可能感染疾病。为什么是这样的，以及我们如何知道“足够多”意味着多少？

通常来说，那些感染者将会把疾病传染给其他人，但是那些痊愈了的就会获得对这种传染病的免疫。所以我们把人们分类成易感人群、受感染人群和去除人群，最后一个指的是那些接受疫苗注射、感染后康复、物理隔离或者已经死亡的人。将这四种情况归为同一类看

似有一些生硬，但残酷的事实是，仅从流行病传播的角度，这四种情况是完全等价的。

为了考察流行病是如何发展的，设S为易感个体数，I为受感染个体数。这两个数字的乘积给出了受感染个体与易感个体可能接触的总数。城市中拥挤的全体居民要比同样多的农村中分散的全体居民更频繁密切地接触，而一个易感个体在一次接触中被感染的概率依赖于疾病的传染性。总的来说，在任意很小的时间段内，一些易感个体被感染的概率形如$\beta \times S \times I$，其中β同时依赖于疾病的传染性和人们混合的情况。

在这个很短的时间段内，任何一个受感染个体可能会被移入去除人群中。因此受感染人群减少一个成员的概率将会与受感染的人数成正比，所以我们采用$\gamma \times I$的形式，其中γ依赖于感染被治愈、隔离或者致死的速度。

我们已经分别得到了受感染人数上升和下降的概率。它们($\beta \times S \times I$和$\gamma \times I$)之间的平衡将决定流行病是否会发生。这十分类似赌博。如果游戏局势对你不利，每一个赌局都会减少而不是增加你的资金。你的资金遵循着随机游走的规律，不可避免地漂移到0。但是如果游戏局势对你有利，例如坏运气在游戏初期没让你破产，此时的随机游走就会使你从所有输钱的趋势中抽离出来，使你的资金远远超过0。要赢得大额赌金的话，游戏对你有利这个条件是必要的，但不是充分的。

在这个流行病的情境中，这就意味着流行病(等同于大笔收益)只会出现在受感染者数目变化是增加远多于减少时。用符号描述，就是$\beta \times S \times I$远大于$\gamma \times I$，这也等同于要求易感个体数S超过比例γ / β，这个比例被称为易感个体数的阈值(threshold)。这就是我们一直寻找

的：即使疾病传入人群中，流行病也只会发生在易感个体数超过这个阈值的情况下。

威廉·科尔马克(William Kermack)和安德森·迈肯德里克(Anderson McKendrick)在1927年发表了这一结果。将易感人群总数控制在这个阈值以下可以阻止流行病的发生，两种明显不同的方法能达到这一效果。第一，接种疫苗来减少易感个体数；第二，找到增大阈值的方法。这个阈值是一个比值，分子增加的时候，阈值就会增加——加快治愈的速度，或者更迅速地将人们隔离——或者让分母减小——也许可以降低疾病的传染性，或者确保人们更少地相互接触，这可以通过暂时关闭学校或者推迟有大批人群聚集的运动赛事进行。我们还可以评估这些对策的收益，以便决定采取什么策略。

相同的原理对处理动物中的流行病也是奏效的。阻止牛群中口蹄疫暴发的第一步就是限制牛群的活动，这是通过降低分母来提升阈值。与此同时也会伴随着大量屠宰(在人类中不可行!)，这样不仅降低了易感个体，也会提升阈值的分子。

上述分析也揭示了为什么我们更希望两次儿童流行病暴发的间隔时间是相对有规律可循的。在一次流行病将易感个体数量减少到阈值以下之后，没有足够免疫的新生儿会逐渐将易感个体数提升到阈值以上，这就为下一次流行病的暴发创造了条件。两次流行病暴发的间隔越长，易感人数就会越多，暴发就会越严重。

了解概率知识可能不会治愈疾病，但可以缓和疾病的影响。

批量测试

一支军队想要确定1000名潜在的新兵中有哪些是对某种疾病易感从而不适宜服役的，可以进行血液检测，但每次检测花费50英镑。用

少于50 000英镑可以完成这项工作吗？

如果易感人群只占相当小的比例，那么答案就是肯定的。选取数字K，收集K名新兵的血液混成一个样本；然后测试这个样本。如果结果是阴性的，那么所有的这些抽血了的新兵就都是健康的，不需要进一步的检测了；否则(检出阳性)这组人里面至少有一个是阳性，所以我们再进行K次检测，每个人进行一次，来确定到底谁是阳性。如果我们足够幸运的话，一次检测就够用了，但我们有可能需要进行K + 1次检测。我们希望检测次数比每个新兵单独检测的次数低。

对一次收集多少个人的样本是最佳选择依赖于检测到阳性的概率。假设这个概率是1%。那么如果我们一次混合10个样本，那么它是一个阳性样本的概率大约是10%，同时这个样本为阴性的概率是90%。所以90%的情况中我们只需要1次检测，但是10%的情况我们需要11次检测。这致使每次需要平均2次检测。混合10个样本将会把检测10名新兵的花费从500英镑减少到100英镑。如果我们将这1000个士兵每组10人地分成100组，然后混合他们的血液样本，我们预期将会节省最初估计花费50 000英镑的80%。

更加精细的计算显示，如果检测出阳性的概率的确是1%，将每组按11人批量检测会比10人稍微好一点，但是其中的差别很小。但是批量检测分组大小的最佳选择对于检测出阳性的概率是非常敏感的。如果我们预期2%的新兵会被检测出阳性，那么每组8人的时候花费最少；在预期的阳性率为5%的时候，每组5人花费最少；而在预期的阳性率为10%的时候，每组4人最合适。(又一次地，使用二项分布可以得到这些结果。)

在第二次世界大战中，这个朴素的想法为美国节省了最初预期检测花费的80%。

航班超额预订

即使必须对那些支付过机票但被赶下飞机的乘客进行补偿，航空公司还是会常规地售出比飞机载客量更多的机票。其中的原因涉及简单的经济学：航空公司完成一次航班的开销几乎不受乘客数目影响，但是每个空座位都会使收益减少。如果预期不是所有订了某次特定航班机票的乘客都会乘机，那么航空公司如何计算出最合适的超额预订数量？

假设飞机上有100个座位，每个座位的费用是100英镑，但是如果因飞机上位置已满而拒绝一名乘客登机需要赔偿其200英镑。航空公司需要对一名预订了机票的乘客实际乘机的概率进行较好的估计。对于前往旅游目的地的包机来说，这个概率将会接近于100%，但是对于那些旅行计划中有更多可变性的乘客来说这个概率会比较低。有关频率的数据将会帮助航空公司估计这些概率。

假定每个预订机票的乘客都独立地有80%的概率会最终登机。如果售出120张票，总的收益就是12 000英镑，而且，虽然平均只有96名乘客会登机，但是有一定的概率，大概15%，超过100名乘客露面，至少有一名乘客留在地面。(这些数字再一次由二项分布给出。)这种情况下航空公司因为超额预订而需支付的赔偿平均为80英镑。多售出5张机票将提升500英镑收益，而平均全部赔偿只提升了275英镑，所以这个策略预计会收益更多。平均收益最多的是售出128张票——比售出125张更多，额外的收益300英镑也已经比平均赔偿295英镑更多了，同时售出129张票会比这稍差一点。

假设某些乘客比另一些乘客更有可能实际登机是更现实的，而且一组一起订票的乘客要么会全部出现，要么全部不出现。这些细节会被容纳到模型中，航空公司会继续售出座位，直到多售出一个座位预计的赔偿超过了额外收益的增加。

排队

概率发展得最好的应用之一是在各种排队中。最初的需求源自试图理解电话线路中的阻塞现象——为表扬丹麦电话工程师艾格纳·埃朗(Agner Erlang)的工作，电话通信流量大小以他的名字命名。排队论为1948年和1949年的柏林空运[3](The Berlin Airlift)的胜利作出了贡献，对排队的系统性的学习在接下来的20年中蓬勃发展。

大卫·肯德尔(David Kendall)引入了一种记号，形如A/B/n，现在普遍作为在顾客逐个地加入队列的时候，对于队列的描述方法。第一个组成部分A指的是两个顾客到达的间隔时间的分布，B指的是为1个顾客提供服务所需时间的分布，n是服务员的数量。

例如，在表达式D/D/3中，“D”是确定性(deterministic)的简写，它意味着完全没有随机性。顾客到达队列末尾的间隔时间是恒定的，所有的服务时间都恰好相等，并且有3名服务员。这种队列在概率的领域中可能并不让人感兴趣，因为其中没有变化性。但若假设有大量的顾客，在很短的时间间隔内他们中的每一个都有很小的概率加入队列，所以顾客以一个总体来说平均的时间间隔加入队列，但是顺序完全随机。这种情况下使用符号M，为了纪念安德雷·马尔可夫，所以M/D/2意味着顾客们随机地到达并且选取两个服务员中的一个，这两个服务员完成工作花费的时间是固定的。

我们想知道队列是如何运作的。最主要的问题是顾客们将会等多久，服务员多么频繁地无所事事，还有为了改善情况我们可以做什么？“服务员”可能是重症监护病床，同时“顾客”就是等待看护的患者。

如果平均每5分钟有一个顾客到来，一共有3名服务员，那么除非平均服务时间少于15分钟，否则就将会产生一个无限长的队列，总体

运营就无法维持了。所以我们必须假设考虑到了服务员数量之后，平均服务时间小于顾客加入队列的时间间隔。这两个平均时间的比值被称为**流量强度**(traffic intensity)，是一个介于0和单位1之间的数值。

在理想情况中，顾客排队时间比较短而且服务员将会一直工作。但是这两种要求正好是矛盾的。以简单的情况为例，有一个服务员而且顾客随机到达。如果流量强度是0.9，计算显示我们预计也许平均有5名顾客在等待，在10%的时间中队伍中没有人。如果流量强度增加到0.98，服务员将会只在2%的时间内空闲，但是平均队伍长度将会升高到25。大多数的顾客将会认为这是个比较差的安排。除非服务员们有更多的“空闲时间”，否则顾客们将会生气或离开放弃这里的服务，或者又生气又离开。

队列的运作不只是依赖于流量强度。在其他条件相同的情况下，服务时间的可变性越强，预期队列将会越长。在只有几个服务员的情况下，顾客加入唯一的一条队伍之后分配给6个服务员(例如火车站)，与顾客可以选择加入某一条队伍之后等待(例如超市)是不一样的。在某些情况下(例如叫救护车)某些顾客可能会有更高的优先级。许多的队伍遵循着“先到者先服务”的规则，但是在不易腐烂的货物存放在架子上等待使用时，可能会遵循“后到者先服务”的规则。某些队伍会汇入其他队伍，服务员们工作速度不同，一群顾客可能会同时到来。明察秋毫的排队论理论家们可以解答几乎所有你能想到的实用模型中的核心问题。

[1] 20世纪40年代，科学家冯·诺伊曼(John von Neumann)与斯塔尼斯拉夫·乌拉姆(Stanislaw Marcin Ulam)等人于美国洛斯阿拉莫斯国家实验室(Los Alamos National Laboratory)为核武器计划工作时，发明了蒙特卡罗方法，此方法的名称来源于乌拉姆的叔叔经常光顾的位于摩纳哥的蒙特卡罗赌场。

[2]　又称21-三体综合征、先天愚型，是一种因为21号染色体的三体现象造成的遗传疾病。

[3]　英美等国为解除冷战初期苏联对西柏林的物资封锁而展开的空运行动。

08 其他应用 Other Appl ications

这一章中我们不关注概率在掷色子、赌场赌博和其他在自然科学中的种种应用。这一章中为了展示概率的普遍存在，我找出了它在法律、社会科学、体育运动和经济学中的应用。

法律事务

尽管20世纪最著名的英国法官之一——丹宁勋爵(Lord Denning)有一个数学学位，但是很少有律师对概率抱有好感。有关概率这门学科有关的措辞在法庭中被自由使用，这应该是令人震惊的。在民事案件(例如诽谤案件)中，陈述“概率是均衡的”就相当于明确地将分割线设在了50%。但是在刑事案件中，只有在陪审团“确认”有罪后他们才能定罪。在数字上没有共识。其中一些人在她们80%确定有罪时定罪，另一些人会在95%或者更高时确定。这些明显是主观概率。而且尽管无论在什么样的罪行判定中都使用相同的措辞，有些人会对相对较小的罪行采取较低的证明阈值。这也就是为什么比起逃票，恶性谋杀案更难以定罪。

假设一名专家证人做证说被告的DNA与在案发现场找到的DNA相符，而后者与随机选取的一个无辜者相符的概率是几百万分之一。陪审员们对这个证词的理解可能存在两个不同的问题。第一个是他们也许会认为证词等价于说在犯罪现场发现的DNA不属于被告的概率是几百万分之一。第二个是他们会将所有这些很小的数字等量齐观，即使一千万分之一与十亿分之一的一百倍是不同的。

第一个问题被称为“检察官谬误”(prosecutor’s fallacy)。很明显这个错误就相当于认为已知DNA相符的情况下无罪的概率，与已知无罪的情况下DNA相符的概率，是相等的。这是个逻辑上的无稽之谈。

已知轮盘赌轮是公正的时候，转到0的概率与转出了0的情况下赌轮公正的概率并不相等。向陪审团提供与犯罪现场DNA相符的市民的数量会避开这个陷阱。在有大约6000万人口的情况下，如果相符的概率是两百万分之一，就会有30人左右相符；如果概率是两千万分之一，大约有3个人相符，不太可能超过半打。但是不要忽略了“随机选取”这句话：与罪犯亲缘越近的人，越可能相符，这个证据对被告不利的程度就越弱。

通过考虑到贝叶斯公式计算证据有效性的方法，可以避免第二个错误。在这个证据呈递之前，你对被告是否有罪有一定的看法。如果证据出现的可能性在有罪时是无罪时的10倍，那么有罪的赔率就会在证据出现之后乘以10；如果证据出现的可能性在无罪的时候是有罪的时候的3倍，那么有罪的赔率就会减少到原来的1/3，以此类推。在存在DNA证据的情况下，假设有罪时证据出现的概率经常是100%，这就让证据的影响变得清晰起来：有罪的赔率就应该乘以那个是“几百万”的具体数字。

随机化回答

一名班主任想要查明高年级学生吸食大麻的比例。直接询问无法得到可信的答案，但是一种称为**随机化回答**(randomized response)的技术是可用的。大意就是负责记录回答的老师并不知道实际上询问的问题是什么，所以大麻使用者就不会担心被认出而可以诚实地回答问题。

80张卡片每张都写着“我吸食大麻”，另外20张都写着“我不吸食大麻”。每张卡片都被装入完全相同的信封中，这100个信封在一个大袋子中被充分混合。学生们应该看到这些操作，所以他们明白袋子中装着两种不同版本的问题和它们的比例。

安吉拉随机选择了一个信封，打开了它并自己看了问题，然后仅仅回答“同意”和“不同意”中的一个。然后她将卡片放入信封，将信封返还到袋子中，并摇晃袋子为下一个学生做准备。

假设1/3的回答都是“同意”。因为学生们都是随机地抽取信封，“同意”来自80%的使用者，与20%的非使用者的诚实回答。几行代数计算显示这与2/9的学生是使用者的情况是一致的。班主任得到了他的答案，没有单独哪个学生被认出来。

或者，把“我不吸食大麻”替换成一个无关的题目，它被“同意”的比例是已知的。如果一个早期的调查估计有一半学生拥有一只宠物，而且没有理由将吸食大麻和拥有宠物联系在一起，在那20张卡片上的问题就可以是“我拥有一只宠物”。然后，如果1/3的回答是“同意”，我们估计7/24的学生是大麻使用者。

得出这些答案的计算过程会在附录中给出。

每次提问时问的是哪个问题的不确定性致使最终估计中存在一些不精确度。装有敏感问题的信封的占比应该尽可能高，但是要足够低以使真正的大麻使用者相信给出诚实的回答不会招致麻烦。在高达95%的卡片上设置敏感问题是不会有效果的。

世界反兴奋剂机构

世界反兴奋剂机构(The World Anti-Doping Agency，WADA)致力于促进体育运动成为一项健康的活动，他们识别使用能提升表现的药物的运动员，并将之排除在比赛之外。但是无论使用什么方法，任何检测程序都可能会给出两种相反的错误：认为一名无罪运动员使用了禁药，和让一名禁药使用者通过检查。

不幸的是，减少两种错误之一的概率的方法经常会造成另一种错误的概率增加。例如，检测测量睾酮(testosterone)与表睾酮(epitestosterone)的比例。正常情况下肌体产生相等数量的这两种物质，但是那些通过注射睾酮以作弊的运动员就会有较高的T/E比。T/E比高于某个特殊的值，比如6比1的运动员，将会被禁止参赛。然而，自然情况下T/E比就会变化：它会随着月经周期(menstrual cycle)变化，如果你得流感了，它也会增加。将T/E比临界值设定得太高则没有药物作弊者会被检出，设置太低则许多无罪的运动员将会被错误地指控。

假设某一次检测出错误的概率是1%。这意味着如果这个运动员是无罪的，那么他们不通过检测的概率是1%；如果他们是禁药使用者，那么他们通过检测的概率也是1%。萨姆没有通过测试：她是无罪的概率是多大？

问题这样描述，那么你几乎会脱口而出认为是“1%”——这个测试每100次出现一次错误，所以如果她没有通过测试，出错的可能性就是1/100。你要克制这种冲动。唯一的正确答案是：“我不知道，这个概率可能是任何值。我们需要知道在全体受试者中使用禁药作弊的比例。”

比如，假设这个比例是1%左右。那么在10 000个运动员中预计有100个禁药作弊者，和9900个无罪者。在检测中，我们预计只有1个禁药使用者通过了测试，其余99个都不能通过测试。但是9900名无罪者中的1%，也就是说99名运动员，也会无法通过测试。在那些无法通过测试的运动员中，一半是无罪者：那么萨姆是无罪者的概率就是50%。

如果禁药作弊者比例不同于1%，结论就会变化。如果它升高，萨姆无罪的概率就会降低，但如果它降低，她无罪的概率就会升高。禁

药使用者的比例越低，这项测试就会越不令人满意，尽管它显然表现出色。

在我们考虑如何检测出机场中的恐怖分子的时候，相同的逻辑也会奏效。无论什么筛选设备都不会是完美的，但是假设一名真正的恐怖分子规避掉这些检测的概率是很低的，比如说1/10 000，同时一名无辜的人被带进小屋进行严密审讯的概率是1/100 000。那么一个被揪住的人有罪的可能性有多大？

我们在不知道准乘客中恐怖分子的比例的时候没法回答这个问题。试作1/1 000 000——考虑到希思罗机场(Heathrow Airport)每年客流量为5000万，这高得恐怖。但是这个数字使我们确信，即使有50名潜在的恐怖分子，他们全被检测出也是极有可能的。

但不幸的是，500名无辜的乘客也将会被扣押！在那些被检测系统拦截的人中，只有少于10%是恐怖分子。而且如果恐怖分子少于50名，被拦截的人确实有罪的概率甚至更低。检测方法想要有效的话，准确性需要更高。

足球比赛结果(1)

在英国，赌足球比赛的结果会引起人们的极大兴趣。无论赌局多么奇异都会有人捧场——赌第一个球的进球时间、所有进球球员所穿球衣的号码之和、比赛中判多少次黄牌和红牌——但最引人注目的是比赛结果是主队获胜、平局、客队获胜这三种中的哪一种。一个理性的下注者会评估这三种结果各自的概率，而他是否下注、下注大小，还有赌本的赔付金额(payout prices)都会基于这些评估。

但是下注者如何推断出他对不同结果的可信度呢？2009年5月，统计学家大卫·斯皮格霍特(David Spiegelhalter)接受BBC广播节目

《 更多或更少》(More or Less)的挑战来分析两天后进行的10场英超联赛。对每一场比赛，他都基于对每一支队伍的进攻能力及其对手的防守能力的考察，给出了每支队伍平均的进球数。例如，一个强势的主队，如阿森纳(Arsenal)在对阵斯托克城(Stoke City)的比赛中估计平均会进球2.1个。

没有任何一支球队会进球2.1个，这个数字只是假想数量的比赛中进球数的平均值。最重要的步骤是评估在单场比赛中进球数为0、1、2、3……的概率，斯皮格霍特使用了泊松分布。多年以来的数据显示，这是一个描述实际进球数在其平均值附近变化情况的好方法。在阿森纳平均进球数为2.1时，没有进球的概率是12%，进1球是26%，进2球是27%，进3球是19%，等等。

斯托克城的数据显示他们的平均进球数是0.67。这相当于没有进球的概率是51%，只进1球是34%，进2球是11%，等等。我们姑且相信每支队伍的进球数都是无关的。所以比赛结果为2比1的概率来自主队进2球的概率乘以客队进1球的概率——在上述情况中为27%× 34%，大约为9%。

在这种方法中，任何可能的比分结果都被估计了。然后主队获胜、平局、客队获胜各自的概率就可以用加法定理计算得到，即将能得到这三种比赛结果的比分的概率分别加在一起。这就得到了阿森纳获胜的概率是72%，斯托克城获胜的概率是10%，平局的概率是18%。比分结果2比0的概率最高，是14%。

别得意! 10场比赛中最高概率的比分才出现2次，而8场比赛的比分都是已被认定为不是最可能的。赌徒已经把钱押在了每个“最可能的结果”和每个“预测的”确切分数上，即使比赛比分结果未揭晓，他也会笑得很开心。

我们不能让阿森纳进行同一场比赛100次来计算他们获胜的频率，那么我们如何调和阿森纳获胜72%的可信度与概率的频率观点呢？回忆我们是如何评价天气预报的准确性的，播报员说明天下雨的概率是30%。只有一个明天，明天要么下雨要么不下雨。然而，我们可以查看她给出30%降雨概率所对应的所有场合，并计算第二天确实下雨了的频率。我们应该根据她播报的整体记录，选择相信或不相信她对明天天气的播报。在足球比赛中，我们也可以对于整个赛季中进行的所有比赛做类似的计算。在这些比赛中(也许共有40场)，对其中一些比赛我们给出了接近72%的概率——我们可以检查“预测的”结果真的发生了的频率是否在72%左右，这可以作为一种验证我们方法的方式。

一个赌徒通过使用这些方法就能够赢钱了吗？赔付金额很大程度上依赖于每种比赛结果被下了多少注，最大的金额通常都会被下注到一支队伍的获胜或者另一支队伍的获胜。忠诚的粉丝们不会倾向于下注到平局上。如果评估得到的平局概率是25%，而且赔付高于3比1，赚钱的机会就在这里。

所以不要以为最好的下注方式就是押在有最高预测概率的比赛结果上!

足球比赛结果(2)

在2010年国际足联世界杯决赛前夕，统计学家伊恩·麦克海尔(Ian McHale)发表了他的计算结果，为32支球队分别指定了赢得世界杯的非零概率。他认为西班牙队最有可能获得世界杯，尽管只有11.6%的可能性；紧随其后的是巴西队，他对其给出的概率是10.3%。

为了计算得到这些结果，麦克海尔使用了与上述计算类似的方法来计算每场比赛的结果。然而，他并未直接计算每场比赛不同结果的概率，他依靠蒙特卡罗模拟。

英格兰队每场比赛的平均进球数为1.5，泊松分布模型给出没有进球的概率是22%，进1球的概率是33%，等等。计算机的随机数生成器选取0、1、2、3……中的某一个值和与其对应的概率，并对英格兰队的对手做相同操作，得到一些模拟比分结果例如2比2平局。对每一场预定的比赛都进行类似的模拟，得到模拟分组表格，然后进行淘汰赛阶段的比赛直到决赛。这个过程重复了100 000次，每一支队伍成为总冠军的模拟数都被记录下来。西班牙“获胜了”11 633次，因此有前述的11.6%这一数字。像往常那样，大数定律是这个方法合理的理由。

那一年西班牙队的确获胜了!麦克海尔的概率是“正确的”吗？我们不知道，如果能进行无限次重复的联赛，也许西班牙队会在其中的65%获胜。但是能证明其方法合理的最好证据就是，庄家们设定初始赔付金额时沿用同样的方法来吸引下注者。

布莱克-斯科尔斯模型

股票市场上的股价会波动，有些时候没有明显的原因。如果今天价格是5英镑，你不知道下个月会是什么价格。然而，你可以购买**期权**(option)——在一个指定的未来时间按照**行使价**(strike price)5.2英镑购入(或者出售)的权利。如果在未来的那个时候，市场价少于5.2英镑，你应该不会行使期权来购买；但是如果市场价高于这个价格，通过行权买入并立即出售你就可以立即获利。相应的论述也适用于认沽期权。那么这些期权的合理价格是什么？

费舍尔·布莱克(Fischer Black)和迈伦·斯科尔斯(Myron Scholes)于1973年解决了这个问题。他们工作的核心是假设股价会随机变化，但是变化方式是一种与高斯分布有关联的特定方式。认购期权与认沽期权的合理价格都依赖于当前价、行权价格、干预时间、现行利率和相关价格的波动性(用一段时间内的标准差来衡量)，但是**不**依赖于平均的价格预期值!

最后一点似乎是令人惊讶的，但这就是现实的运作规律。这也是非常有用的，这意味着我们避免了任何因估计价格趋势而引入的不确定性。如果你想要知道一些特定期权的合理价格，有免费的软件可用——只需要在你最喜欢的搜索引擎中输入“布莱克-斯科尔斯”。给定当前价格和行权价格，如果时间间隔变长，利率变高，或者价格波动性升高，一份认购期权的合理价格会随之增加。

刚才那句话与你的直觉相符吗？第一个情况似乎的确是合理的，因为你等待的时间越长，相关价格上涨的概率就越大，但是另外两个情况就更加微妙了。波动性的增高会使得价格上涨的概率更高，但同时也会让价格**下跌**的概率更高——但是前者的效应更强一些。

通过考察这一年中约250个交易日内价格的变化，可以衡量得到波动性。为了使估计是可靠的，应该向模型提供足够的数据，但是也不能向过去延伸得太远，因为那些数据与现在的情况不相关。对于波动性的不良估计会导致对期权合理价格估计的不合理。

只有在假设合理的情况下一个模型才是有效的。如图7所示，认为价格波动符合高斯分布就意味着严重事件的概率是非常小的，例如价格下降了3或4个标准差。当这些事件的真实概率被严重低估的时候，模型就是无效的，而模型导出的结论就根本没有合理的基础。在第4章中提到过的极端值分布可以用来解决这个问题。

投资分配

A公司与B公司都预计会盈利。在低利率的情况下，A预计会收益20%， B预计会收益40%；在高利率的情况下，情况反转——A会盈利40%， B是20%。假设尼克是一名规避风险的投资者，而玛丽喜欢冒险。

如果高利率和低利率可以被认为是同等可能的，两个公司看起来都很吸引人，收益平均都是30%。依据两人对风险的态度，尼克会将资金分成两等份投给两家公司，而且一定会获得30%的收益，无论实际利率是高是低；玛丽会把所有钱投到其中一家公司，并希望会获得40%的收益，但也得接受只获得20%收益的结果。

假设情境中C公司取代了B公司，其会在低利率的时候有10%的收益，或者在高利率的时候有50%的收益——同样的平均为30%，这一点与B公司一样。但是将A与C组合对两名投资者都没有意义：尼克只钟意于A，玛丽将所有的钱投入C。

两种情况的本质区别是，A与B的收益是**负相关**(negatively correlated)的——在一个比较高的情况下，另一个比较低；但是A与C的收益是**正相关**(positively correlated)的——它们一同变好或者变坏。“相关性”(Correlation)以一个在-1(完全负相关)和1(完全正相关)之间的值来衡量。如果每家公司的资产波动变化都独立于另一家，它们的相关性就是0。

规避风险的投资人倾向于分散持股，所以任何损失都会被其他地方的收益所弥补。它们希望持有负相关的资产。但是其中有一个无法避免的逻辑陷阱：如果X与Y负相关，Y与Z负相关，那么X与Z就似乎是正相关了!

然而，还有一线希望。所罗门·博纳(Salomon Bochner)的数学工作结果证明了在很大的投资组合中，每一对投资都是负相关的确是**有可能的**。但是投资可选的数目越大，两两负相关的情形就越难以达到。

09 有趣而棘手的问题 Curiosities and Dilemmas

在这本书的开头，我曾经说过一些概率问题第一眼看上去有违常识。随着故事的逐渐呈现，相关的例子已经讲述过了。这里呈现一些直觉会产生误导的情况，但是足够小心的话，这些表面上的矛盾都是可以解释的。概率这门学科已经完全不含有真实的悖论了。

但是即使概率知识能够帮助我们做出合理决定，我们也许仍然会发现，就算是考虑某个特定事情的概率也可能会遇到棘手的问题。

帕隆多悖论

格雷厄姆·格林(Graham Greene)的小说《败者为王》(Loser Takes All)是一本好书，但是它基于一个错误的前提：有一些巧妙的基于数学的下注组合方法能让玩家占有优势而不是庄家。但数学家们已经证明，在每一场赌局单独来看都是对庄家有利的时候，无论怎样组合都不会扭转局面对玩家有利。对不起啦，朋友们。

胡安·帕隆多(Juan Parrondo)告诉你，你务必十分明确地阐述如下一般论断：在所有赌局都对一方有利的时候，无论何种情况我们都不可能找到一种组合让另一方有优势。我在这里描述一种他的思想的变体，其来自迪恩·阿斯图米安(Dean Astumian)，他描述了一种依托于一张画有5个格子的纸板的简单游戏，如图11。(这不是一个**认真的**游戏。它的存在只是为了阐述上述观点)

输	左	起始	右	赢

图11　阿斯图米安的游戏纸板

你需要一种生成可能性为1%的随机事件的方式：也许是一个装有99个白球和1个黑球的袋子，或者一个会等可能地停在100个小格中的转盘。游戏开始时，在标有“起始”的格子中放一个标记物。每一次移动都将标记向左或者向右移动一格，如果在到达赢之前没到过输就算胜利。

一共有两组基本规则，我们叫它们A和B。在规则A中，你总是会从起始移动到左；你一定会从右移动到赢；你会在左使用转盘，有1%的概率移动到输，99%的概率移动到起始。在规则B中，你在起始使用转盘，有99%的概率移动到右，1%的概率移动到左；在右，你总是从右移动到起始；你在左时，情况和规则A中一样——转盘给出1%的概率移动到输，99%的概率移动到起始。

对这个游戏的分析是很简单的，在规则A中，没有规则允许移动到右；你在起始和左之间往复移动，直到随机概率使你从左移动到输。在规则B中，你经常在起始和右之间移动，而偶尔移动到左。最终，在其中一次经过左的时候，随机概率使你移动到了输。在这两种游戏中，移动到赢的概率都是0。

在一个新的游戏规则C中，你需要一枚公正的硬币。在每次移动前，掷一次硬币：如果正面朝上，使用规则A；如果背面朝上，使用规则B。

结果是你在规则C中取胜的概率超过了98%！很容易就可以阐明为什么游戏对你十分有利：如果你到了左，你非常有可能安全回到起始。从起始，有一半可能你使用规则B，有99%的概率到达右；而在右，你有一半概率使用规则A，必然会取胜。

遵循着规则A和B，你**一定**会输：在两种规则之间交替，你几乎每次都会赢! **悖论中**构筑一个排除了上述例子的数学命题需要非常严谨的语言描述，这也就说明了格林的结论其实摇摇欲坠。

2 + 2 = 4，还是2 + 2 = 6？

假设我们用一个公正的硬币来实现伯努利试验，也就是说，每次投掷硬币结果都是独立的，且正面(Heads)反面(Tails)是等可能的。典型的结果是HHTHTTTHT……要掷出H的平均等待投掷次数是2；但是要掷出HT，或者HH的平均等待投掷次数是多少呢？

直觉上讲，答案是4，因为我们预计会等2次投掷来得到第一个标志H，等2次投掷得到另一个标志T。我们等待HT的平均投掷次数的确是4，但对于HH**不是**这样。为了看到这种样式，平均投掷次数是6!

有这种不同是因为，要得到HT，认为我们预计得等2次才得到H是正确的，而要再等2次才得到T从而完成这个样式。2加2等于4。但是对于HH，在我们得到第一个H的时候，下一次投掷有一半的可能性会掷出T，我们就得重新开始了——之前得到H的所有次数就都浪费了。得到正确答案的计算过程在附录中。

H与T两者中一个先出现是等可能的；那在HH与HT之间呢？再一次地一个比另一个出现得早是等可能的，因为我们必须等到第一个H出现，之后下一次投掷决定了最终结果。然而，在HH与TH之中，后者首先出现的可能性是前者的3倍! 原因很简单：序列由HH开始的可能性是1/4，但是除非这样，那么不可避免地TH首先出现(思考一下这是为什么)。

彭尼赌局游戏(Penney-ante)就是基于上述的观点。你请你的对手选择8个可能的长度为3的一组结果中的任何一个，比如HHT，或者THT

等，它们都可能会是连续3次投掷公正硬币 的结果。之后你选择一个不同的结果，一个中立的人重复投掷硬币，选择了首先出现的结果的那个人获胜。

尽管表面上看你大方地允许你的对手先选择，但是这个游戏是对你有利的——如果你知道你应该做什么的话。无论她选择了什么，你都可以选择一个有至少2/3可能性比她的样式先出现的样式! 获胜的秘诀在附录中。

给我点暗示……

1. 三张形状大小完全相同的双面卡片装入一个袋子中。其中一张两面都是蓝色，另外一张两面都是粉色，最后一张一面是粉色，另一面是蓝色。随机选择一张卡片，可以看到它的一面是粉色。另一面更有可能是蓝色还是粉色呢？或者说可能性相等？问题交给你吧——下面有回答。

2. 细致的计算表明，从洗好的牌堆中分发出来的一副13张的桥牌，有26%的可能性包含2张或者更多张A。你给露西发牌。对于问题：“你的手牌中至少有一张A吗？”她的回答是“是的”。在另一个情形中，你给蒂娜发牌，并问：“你的手牌中有黑桃A吗？”她的回答也是“是的”。哪一副手牌更有可能包含2张或者更多张A？或者说可能性相等？答案在下面。

3. 假设1000名男性和1000名女性都有令人满意的资质，但有480名男性和仅240名女性获得大学录取资格。这是不是明确的性别歧视——男性被录取的概率是女性的2倍？

答案是什么呢？在粉/蓝卡片问题中，看到粉色就明确地排除了两面均蓝色卡片的情况。所有3张卡片都是完全相同的，只有两张剩下

了，粉/粉和粉/蓝。这些卡片中的一张，背面是蓝色的，而另一张，背面是粉色的。似乎粉色和蓝色是同样可能的。

这个推理过程是草率的：粉色的可能性是蓝色的两倍，你可以通过重复进行十几次这个试验来验证这件事。更好的理解是，注意到这些卡片中有3个粉色面，所有这些面都等可能地被看到。但是只有一个粉色面的背面是蓝色——而有两个粉色面的背面是粉色(你可以使用贝叶斯公式，但是那就是杀鸡用牛刀了)。

一名贫困的研究生，瓦伦·韦弗(Warren Weaver)，同时也是**信息论**(Information Theory)的创建者之一，就曾经不断地和其他学生玩这个游戏并赢钱，教育了要他们了解概率的效用。

在牌组问题中，我们知道两个情况中手牌里都有至少一张A，许多人都会认为蒂娜和露西有2张或者更多的A是等可能的——所有的A都是等可能被抽到的，所以为什么蒂娜承认她有特定的一张黑桃A就会带来不同呢？请你丢掉这些想法，来做正确的计算。

对于露西来说，在有至少一张A的手牌中，我们可以算出有2张或者更多A的比例——大约37%。对于蒂娜，除了黑桃A，她还有另外12张手牌，从剩余的51张中随机选取。手牌中包括另一张A的可能性是56%：蒂娜远比露西更有可能持有2张或者更多A。

你的怀疑心告诉你第三个问题正确的答案是“否”。假设在英语系，950名女性申请者中的20%，以及50名男性申请者中的10%被录取了；在商学院，所有的50名女性得到了录取资格，但是950名男性中只有一半被录取。求和得到：240名女性和480名男性被成功录取，但是，**在每一个院系**，女性的录取率都是男性的2倍。有歧视的话也是针对男性的，而不是女性!

确实，在真实世界中，伯克利[1]研究生院的几千名申 请者中，44%的男性被录取，而只有35%的女性被录取。然而，当申 请数据被分配给不同的院系的时候，男性和女性的录取比例就几乎没有差距了。但是不同的院系的录取率的确不同，而那些对两种性别都只录取很小比例的院系，女性申 请者最多。

这个反直觉的结果是**辛普森悖论**(Simpson’s Paradox)的一个例子。它展示了相较于对绝对数字的操作，对比例的操作是很危险的，这种情况到处都会发生。

这一切都不只是好玩。除非你知道数字的真正含义，不然你没有正当理由来说你会用数字。

你真的想知道吗?

我曾经说概率是在不确定性中做决定的关键，我也不会收回我说的话。但更加精确地理解概率或者在新的情境中理解概率都会带来一些令人不舒服的难题。

现在，个人可以对自己整个遗传密码进行测序，但是诺贝尔奖获得者詹姆斯·沃森(James Watson)和哈佛大学心理学家史蒂芬·平克(Steven Pinker)都选择不去知道他们携带的一种被叫作APOE的基因的版本。有一个epsilon4版本的这种基因会让患上阿尔茨海默病[2]的概率上升4倍，而有两个这种基因会让概率升高20倍。(矛盾的是，有这种epsilon4基因也与一个人某些年轻时的益处相关。)另一名诺贝尔奖获得者，克雷格·文特尔(Craig Venter)，知道他的确有一个epsilon4基因。一家研究实验室有**从不**向志愿者透露其APOE基因情况的政策，理由是基于现在人类掌握的知识，没有治疗可以减轻其带来的消极影响。

但是一些商业公司也许会对你的APOE基因的情况(实际上是你的全部基因组)**非常**感兴趣。如果你的基因组成暗示着你早逝的概率非常高，它们也许会愿意大幅提高养老金——但是也可能会要求更高的医疗保险费。拥有某个人全部基因信息的公司可能会“提供”定制服务，完全按照客户的预期寿命量身定制。

约翰和汤姆都是65岁，每个人都会花费15 000英镑来购买养老金；比如说正常的剩余预期寿命是15年，但是约翰的基因暗示着长10年的寿命，而汤姆是缩短10年。不考虑基因情况的A公司对两人每年都提供同样1000美元的养老金。但是B公司考虑到了基因信息，向汤姆每年提供3000美元的养老金，但只向约翰每年提供600美元。

回想那条格言：从长远来看，平均统领一切。两个人都会接受更高的出价，所以A公司将会给像约翰这样的人支付25 000美元，每次都会损失10 000美元；同时B公司将会向汤姆和他这个类型的人支付15 000美元，所以收支平衡。A公司将会倒闭。而B公司会生存下来。

如果只有像B那样的公司才能生存下来，那么我们可以预见到那时会有许多不幸的人，他们要么是根本无法支付医疗或者旅游保险，要么因为被提前告知储蓄不足，导致退休计划被严重扰乱。

律师在诘问中应该只去问那些他们已经知道答案的问题。在你想要对自己的基因组进行测序的时候，要确保自己对你可能得到的消息有充分的准备。考虑一下人生中的所有阶段：在孩子出生时的基因组情况的打印件也许会带来晴天霹雳；想结婚时，你和与你订婚的人是否应该去了解你们孩子患有严重疾病的可能性？你的老板是否应该有权利因为你患某种疾病的风险较高而拒绝你的晋升？高级公职人员，比如总统或者首相的候选人，是否应该公开他们的基因组信息，以便投票者进一步了解任何基因层面的不稳定因素？

随机选择一名英国女性，其罹患乳腺癌的可能性是12%。但是如果她继承了特定的被称为BRCA1或者BRCA2基因的突变，这个概率就会升到60%。一名有3个孩子的母亲，在一名姐妹有这种突变的情况下，接受医学检测，而如果她接受了检测而且收到坏消息，她的女儿(如果有女儿的话)应该在什么年龄被告知她们每个人都有50%的概率继承了这种突变？

无论你在这些令人不舒服的情形中感受到了什么，要记得这只是“概率”而不是“事实”。如果艾玛有这种突变的概率是10%，而菲奥娜的概率是60%，结果也有可能是艾玛患有乳腺癌而菲奥娜没有。如果她们知道自己有这种突变的概率，她们也只能按照自己的方式来处理这件事情。在此重复一下决策论的核心信条：合理的决定是能最大化结果的平均效用。你永远不能确定自己采取的行动会带来最好的结果，但是你已经充分利用了你所拥有的信息。你不能要求更多了。

全书完

[1] 此处应指加利福尼亚大学伯克利分校(University of California, Berkley)。

[2] 阿尔茨海默病(Alzheimer’s disease)，俗称早老性痴呆症，是一种发病进程缓慢、随时间不断恶化的神经退行性疾病。

附录 书中提出的问题的答案

第2章：对于有8或者10个面的色子，掷出偶数的概率是1/2，掷出6的倍数的概率分别是1/8和1/10。掷出3的倍数的概率分别是1/4和3/10，对8面色子来说，掷出偶数和掷出3的倍数这两个事件是独立的，而对于10面色子不是这样。

第6章：在“金钱本色”难题中，你首先应该选择积攒下12 000英镑。如果成功了，你下一次就会希望存下3 000英镑；如果失败了，你下一次就必须尝试存下15 000英镑了。你总体的获胜概率是(6/12)×(10/11) + (6/12)× (4/11) = 84/132 = 7/11，比其他的策略都高。

在有46张相似扑克牌的游戏中，46次重复抽牌中有9次会是黑桃，你会净入账50块；在其余的37次中，你会输掉额外的赌金。如果你多押12块或者更少，长远地看你将会有盈利，但是如果你多押13块或者更多，预期你会输的比赢的多。

第7章：流产的概率是1/m，生出患有唐氏综合征的婴儿的概率是1/n， m和n都很大而且m＞n。只需要看是否有y＞x+n/m。

假设安妮的姐姐或妹妹西莉亚有n个儿子，全都健康。如果贝蒂不是一名携带者，西莉亚的所有孩子都一定是健康的；如果贝蒂是一名携带者，那么西莉亚不是携带者(因此她的儿子们都健康)的概率是1/2，而西莉亚是一名携带者，但是尽管如此她的儿子们都健康的概率是$(1/2)^{n+1}$。所以在有安妮的侄子们都健康这一信息的情况下，贝蒂是一名携带者的后验赔率，就是先验赔率(即利用安妮的兄弟都健康这一信息计算得到的值)乘以后面这些来自安妮姐妹的概率的和[1/2 +

$(1/2)^{n+1}$]。这个值就转化为贝蒂是一名携带者的概率，将其减半就是安妮是一名携带者的概率。

第8章：在随机化回答的例子中，假设吸食者占比为x，而且所有的答案都是诚实的。那么，在第一种情况中，全部回答“同意”的比例是0.8 x + 0.2 (1 - x)；令其等于1/3，解方程得x = 2/9。在第二种情况中，回答“同意”的比例是0.8 x + 0.2/2；令其等于1/3，得到x = 7/24。

第9章：设x为掷出HH的平均次数。为了得到第一个H，平均需要等2次。在这之后，我们至少需要额外的1次投掷；而即使这样，有一半的可能我们需要从头再来。所以x = 2 + (x/2)，即x = 6。

在彭尼赌局游戏中，如果你的对手选择了HHH，你应该选择THH，而你获胜的概率是7/8；如果她选择了HHT，你再一次地选择THH，获胜可能性为3/4；如果她选择HTH，你选HHT，如果她选择THH，你选TTH——在这两种情况中，你有2/3的可能性获胜。利用对称性就可以得到应对TTT、TTH、THT和HTT的最佳选择。

术语对照

A

absolute risk 绝对风险

Additional Law 加法定理

airline overbooking 航班超额预订

Alzheimer’s disease 阿尔茨海默病

amniocentesis 羊膜穿刺术

Astumian’s game 阿斯图米安的游戏

axiom 公理

B

Balls in a bag/urn 袋子/罐子中的球

batch testing 批量测试

Bayes’s Rule 贝叶斯公式

Bayes, Thomas 托马斯·贝叶斯

belief, degree of 可信度

Bernoulli family 伯努利家族

Bernoulli Society 伯努利学会

Bernoulli trials　伯努利试验

beta distribution　贝塔分布

bets　赌注

binomial distribution　二项分布

birthdays　生日

births　出生

blackjack　二十一点

Black-Scholes　布莱克-斯科尔斯模型

Bletchley Park　布莱切利园

Bochner, Salomon　所罗门·博纳

Borel-Cantelli Lemmas　波莱尔-坎泰利引理

Borel, émile　埃米尔·博雷尔

Box, George　乔治·博克斯

breast cancer　乳腺癌

bridge　桥牌

Brownian motion　布朗运动

Butler, Bishop Joseph　主教乔瑟夫·巴特勒

C

card dealing　分牌

Central Limit Theorem　中心极限定理

certainty　确定性

Chebychev, Pafnuty　巴夫尼提·切比雪夫

classical approach　古典方法　参见coin-throwing

objective approach　掷硬币

Colour of Money　金钱本色

computers, infiuence of　计算机的影响

conditional probability　条件概率

continuous distribution　连续分布

correlation　相关性

cricket　板球

criminal trials　刑事审判

D

Deal or No Deal　交易与否

de Finetti, Bruno　布鲁诺·德·菲内蒂

de Moivre, Albraham 亚伯拉罕·棣莫弗

density 密度 参见probability density

dice 色子

diffusion 扩散

dilemmas 难题

discrete distributions 离散分布

disjoint 不相容的

distribution 分布

Doob, Joseph 乔·杜布

E

eggs鸡蛋

Einstein, Albert 阿尔伯特·爱因斯坦

engineering students 工科生

epidemics 流行病

Erlang, Agner 艾格纳·埃朗

errors 误差

expected utility 预期效用

expected value　预期值　参见mean value

exponential distribution　指数分布

extreme-value distributions　极端值分布

F

fair price　公平价格

Feller, William　威廉•费勒

fencing, Olympic　奥运会上的击剑项目

Fermat, Pierre de　威廉•费勒

Florentine gamblers　佛罗伦萨的赌徒们

fiying bombs　飞行炸弹

football results　足球比赛结果

Fréchet distribution　弗雷歇分布

frequencies　频率

frequency approach　频率方法

G

Galileo, Galilei　伽利略•伽利莱

games, recreational　休闲游戏

Gauss, Carl Friedrich　卡尔·弗里德里希·高斯

Gaussian distribution　高斯分布

genetics　基因

geometric distribution　几何分布

Gnedenko, Boris　鲍里斯·格涅坚科

Golden Balls　黄金球

Gorbachev, Mikhail[1]　米哈伊尔·戈尔巴乔夫

Gore, Al　阿尔·戈尔

Griffin, Peter　皮特·格里芬

Gumbel distribution　冈贝尔分布

H

Hand, David　大卫·汉德

haemophilia　血友病

herd immunity　群体免疫

horse kicks, death from　被马踢死

horse racing　赛马

I

impossibility　不可能

independence　独立性

insurance　保险

inverse probability　逆概率

Iterated Logarithm, Law of 《 重对数率》

K

Kelly, John　约翰·凯利

Kendall, David　大卫·肯德尔

Kermack, William　威廉·科尔马克

Khinchin, Alexander　亚历山大·欣钦

Kolmogorov, Andrey　安德鲁·柯尔莫哥洛夫

L

Laplace, Pierre-Simon　皮埃尔-西蒙·拉普拉斯

Large Numbers, Law of　大数定律

Likelihood Ratio　似然比

Longcor, Willard　威拉德·朗克尔

lottery　彩票

M

Malawi, child deaths 马拉维儿童死亡

Markov, Andrey 安德雷·马尔可夫

martingale 鞅

McHale, Ian 伊恩·麦克海尔

McKendrick, Anderson 安德森·迈肯德里克

mean value 平均值

misprints 印刷错误

misunderstandings 误解

Monte Carlo simulation 蒙特卡罗模拟

Morse code 摩尔斯电码

Mosteller, Frederick 弗雷德里克·莫斯特勒

Multiplication Law 乘法定理

mutually exclusive 相互排斥 参见disjoint

N

Nader, Ralph 拉尔夫·纳德

Newton, Isaac 艾萨克·牛顿

Nobel prize 诺贝尔奖

normal distribution　正态分布　参见Gaussian distribution

Number Needed to Treat　需治疗人数

O

objective approach　客观方法

odds　赔率

opinion poll　民意测验

option　期权

Oscars　奥斯卡

P

pairwise disjoint　两两不相交

Parrondo’s paradox　帕隆多悖论

Pascal, Blaise　布莱兹·帕斯卡

payout price　赔付金额

Penney-ante　彭尼赌局游戏

personal probability　个人概率　参见 subjective probability

Poincaré, Henri　亨利·庞加莱

points, problem of　点数分配问题

Poisson distribution　泊松分布

Poisson, Siméon Denis　西莫恩·德尼·泊松

poker　扑克

polymer chemistry　聚合物化学

posterior odds　后验赔率

prior odds　先验赔率

Prisoners’ Dilemma　囚徒困境

probability density　概率密度

prosecutor’s fallacy　检察官谬误

pseudo-random sequence　伪随机数序列

Q

queues　队列

R

radioactivity　放射性

random choice　随机选择

randomized response　随机化回答

random numbers　随机数

random walk 随机游走

relative risk 相对风险

Restricted Choice, Principle of 限制选择原则

Richard III 理查三世

risk 风险

roulette 轮盘赌

Rutherford, Ernest 欧内斯特·卢瑟福

S

safety 安全

sex discrimination 性别歧视

share portfolios 投资分配

Simpson’s paradox 辛普森悖论

soccer, World Cup 足球世界杯

Spiegelhalter, David 大卫·斯皮格霍特

square root 平方根

standard deviation 标准差

Star Trek 《星际迷航》

Strong Law 强大数定律

subjective probability 主观概率

sums 和

survival analysis 生存分析

symmetry 对称性

T

tennis 网球

terrorist detection 检测恐怖分子

Thorp, Edward 爱德华·索普

tiny probabilities 微小概率

traffic intensity 流量强度

trick questions 轨迹问题

TV game shows 电视游戏

U

uniform distribution 均匀分布

utility 效用

V

variability 变化

variance　方差

volatility　波动性

W

WADA　世界反兴奋剂机构

weather　天气

Weaver, Warren　瓦伦·韦弗

Weibull distribution　韦布尔分布

Weldon, Raphael　拉斐尔·韦尔登

Z

zero probability　零概率

[1]　原文误作“Gorbachev, Mikhael”

读客

激发个人成长

多年以来，千千万万有经验的读者，都会定期查看熊猫君家的最新书目，挑选满足自己成长需求的新书。

读客图书以“激发个人成长”为使命，在以下三个方面为您精选优质图书：

1. 精神成长

熊猫君家精彩绝伦的小说文库和人文类图书，帮助你成为永远充满梦想、勇气和爱的人！

2. 知识结构成长

熊猫君家的历史类、社科类图书，帮助你了解从宇宙诞生、文明演变直至今日世界之形成的方方面面。

3. 工作技能成长

熊猫君家的经管类、家教类图书，指引你更好地工作、更有效率地生活，减少人生中的烦恼。

每一本读客图书都轻松好读，精彩绝伦，充满无穷阅读乐趣！

认准读客熊猫

读客所有图书，在书脊、腰封、封底和前后勒口
都有“读客熊猫”标志。

两步帮你快速找到读客图书

1. 找读客熊猫

2. 找黑白格子

马上扫二维码，关注“熊猫君”
和千万读者一起成长吧！

John Haigh

PROBABILITY

A Very Short Introduction

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

© John Haigh 2012

The moral rights of the author have been asserted
Database right Oxford University Press (maker)

First published 2012

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organization. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above

You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Library of Congress Control Number: 2012930333

Typeset by SPI Publisher Services, Pondicherry, India
Printed in Great Britain by
Ashford Colour Press Ltd, Gosport, Hampshire

ISBN: 978–0–19–958848–0

1 3 5 7 9 10 8 6 4 2

Contents

Acknowledgements

I thank Leighton Vaughan Williams and Mike Smith for helpful information on horse racing odds, Les Miller for the results of his simulation of roulette spins, Ian McHale for his detailed explanation of how to estimate the winning chances of the 32 teams in the finals of soccer's World Cup, and Derek Robinson for his help with the diagrams. I am particularly grateful to an anonymous referee, whose acute and constructive comments have helped me remove ambiguities, order the material more coherently, and seek to draw out the central ideas of the subject.

I beg forgiveness from those others who recognize themselves as the unacknowledged origin of the information, opinions, or anecdotes in this book, but I have relied on too many sources to give individual credit every time. This is not an academic treatise, with every assertion traceable to its roots, but an attempt to sharpen the reader's appreciation of what the subject of probability is about, how it has developed, and to where it might be applied.

All errors and imperfections that remain are mine alone.

John Haigh

August 2011

List of illustrations

1 Relative frequencies of Heads in 20 throws 32

2 Times of emission of alpha particles 36

3 Illustration of the Strong Law of Large Numbers 39

4 Some common discrete distributions 45

5 Area represents probability 47

6 A continuous distribution 48

7 Gaussian distributions 50

8 Beta family of distributions 65

9 The general shape of a utility curve 67

10 Family relationships 88

11 The board for Astumian’s game 110

Chapter 1 Fundamentals

The scope of probability

Probability is the formalization of the study of the notion of uncertainty. The effects of blind chance are apparent everywhere. Biologically, we are all a random mixture of the genes of our parents. Catastrophes, like oil spills, volcano eruptions, tsunamis, or earthquakes, and happier events such as winning lottery prizes, randomly and dramatically change peoples' lives.

Many people have a good intuitive understanding of probability. But this understanding can go astray when you have an initial idea about the likelihood of something, but then some new fact, whose relevance is not wholly apparent, is revealed. There are indeed a few notorious 'trick questions', about birthdays, or families with two children, or television game shows with three choices, that seem to have been designed to persuade you that the subject defies common sense. It does not. So long as any hidden assumptions in these questions are flushed out, and taken account of, sensible answers emerge. But probability does require clear thought processes.

The development of its ideas and methods has been driven by its wide applicability. The D-Day invasion of Normandy went ahead in June 1944 only because the probability of favourable weather was deemed sufficiently high. Engineers in

the Netherlands must take account of the chances of severe floods when they build the dykes that protect their country from the sea. Is a new medical treatment more likely than present methods to enable a patient to survive for five years? How much you pay to insure your life, car, house, or possessions depends on the chances of an early claim being made. Most decisions you make - what to study at school, who to select as a life partner, where to live, which career to follow - are made under conditions of uncertainty. As Pierre-Simon Laplace wrote in 1814:

> …the most important questions in life are, for the most part, only problems in probability.

Whenever the phrase 'the probability is…' appears, some assumptions (that may inadvertently have been omitted) are being made. If those assumptions are unwarranted, little reliance should be placed on the claim. I hope that, in this book, these assumptions are clear, either implicitly or explicitly. Before we look at how probability statements can be interpreted, we will describe different ways in which they may arise.

The objective view

The classical , or objective , view of probability is that often used during games of chance, such as rolling dice, or spinning roulette wheels. There is some list of outcomes: then, either from considerations of symmetry, or because we can find no good reason for one of them to occur rather than

another, we take them all as equally likely. So we just count the number of outcomes, and give them all the same probability. Then the probability of any event in the experiment is taken as the proportion of outcomes that favour it.

For example, when a coin is thrown twice, the four possible Head/Tail outcomes are HH, HT, TH, TT. With a fair coin, H or T will be equally likely each time, so none of those four outcomes should be more or less likely than any of the others, each should have probability 1/4. Three of them contain Heads at least once, so the probability of the event that Heads appears at all is 3/4.

There are 1,326 ways of dealing a hand of two cards. (Take my word for it.) If the deck has been well shuffled, we take all these hands as equally likely. And 64 of them consist of an Ace and a 'ten-card' (i.e. Ten, Jack, Queen, or King), so we conclude that the probability of being dealt such a hand – 'Blackjack' – is 64/1326, just under 5%.

So far as probability considerations are concerned, both these examples could be reformulated in terms of choosing one ball from a bag of identical balls. The first bag would have four balls, three of them Red, the second 1,326 balls, 64 of them Red. Indeed, every example in this objective approach to probability is essentially identical to some problem about selecting one ball from some bag or urn (which perhaps

explains the plethora of such exercises in student textbooks).

I emphasize that it is not enough to count the number of possible outcomes, and how many of them favour the event in question. There must also be no cogent reason for any outcome to be more or less likely than some other. Otherwise, you could fall into the trap of believing that your chance of winning the Jackpot in a Lottery is 50%, on the grounds that there are just two alternatives, either you win or you do not!

Experimental evidence - frequencies

We hope that dice used in household games like Monopoly, or casino games like Craps, will show each of their six faces equally often. But if a die is made from non-uniform material, or its width, breadth, and height differ, it is not sensible to assume that all outcomes are equally likely. Over a series of throws made under the same conditions, the frequency of any face will fluctuate, but will eventually settle down close to some particular value.You do not find that 20% of the first thousand throws are Sixes, and then the proportion among the next thousand throws leaps up to 60%. In these repeatable experiments, the outcomes may not be equally likely, but each of them has a propensity to occur at some characteristic frequency, and a frequentist takes this value as the probability of that outcome.

Perhaps we get 170 Sixes in the first thousand throws of an imperfect die, then 181 Sixes in the next thousand, and so on. We can never deduce an exact value for the probability of a Six from these experiments, but the data lead to estimates, and the more data that are collected, the better we expect the estimate to be. The fact that we cannot know the exact probability does not deny its existence. If I draw one card from a well-shuffled pack, there seems no reason for one suit to be favoured over any other. Each suit would have objective probability of 1/4. And if I return the card, reshuffle, and perform this task one hundred times, I expect each suit to arise about equally often, in this case about twenty-five times. Similarly, with ordinary dice where all six outcomes are intended to be equally likely, the chance of a Five on any throw is objectively taken as one-sixth: and over six hundred throws, we expect a Five on about one hundred occasions.

When experiments with equally likely outcomes are repeated often, the relative frequency of any particular outcome is expected to be a close match to its probability, as calculated objectively. A fair coin seldom gives exactly fifty Heads in one hundred throws, but intuition does not tell you how close to that ideal you should reasonably expect. Frequency ideas are applied more widely than to repetitions of the same experiment under identical conditions. Will some imminent birth be male or female? With no specific information about the family in question, turn to data gathered from many countries and cultures over a long

period. There is a consistent pattern that, for every 49 female births, there are 51 males. On the basis that there is nothing to pick out this birth from all others that are taking place, a frequentist will put the probability of a boy at 51%.

Some experiments on a heroic scale have been conducted. In 1894, the zoologist Raphael Weldon reported the results of more than twenty-six thousand throws of a set of a dozen dice. His data were not consistent with the idea that all six faces were equally likely, as the numbers five and six occurred rather too often. His dice had small holes drilled in each face to identify its score, and the faces for five and six are opposite two and one respectively. The centres of gravity of these dice will be closer to the faces with small numbers, giving a plausible explanation for the observed excess.

About seventy years later, Willard Longcor, a meticulous man with time on his hands, offered his services to top Harvard statistician Frederick Mosteller. Under Mosteller’s guidance, Longcor collected over two hundred dice, and threw each of them twenty thousand times, recording the outcome simply as even or odd - over four million data values. To make the conditions as near as possible identical, he used a carpeted desk-top, with a raised step to bounce the dice off. For cheap dice like those used by Weldon, there was a small but distinct bias towards too many even numbers - again, not totally unexpected because of the drilling. However, with the

high quality precision dice as used in Las Vegas casinos, where the pips are either lightly painted or are extremely thin discs, no such bias was found. Frequencies with those dice were consistent with the classical view of equally likely outcomes.

Blackjack expert Peter Griffin noted wryly that, for a sequence of 1,820 hands he played in Las Vegas, the dealer's upcard was either a Ten-card or an Ace on 770 occasions. The objective chance of receiving one of those favourable cards is 5/13, so Griffin wondered whether or not he had been cheated - random chance would give the dealer these good cards only 700 times on average.

In 2002/3, 6,202 children under five years old were admitted with suspected pneumonia to hospitals in Malawi, and 523 died, a fatality rate of 8.4%. Provided that there were no special circumstances making this period atypical, a frequentist would conclude that the probability of death when a young Malawi child catches pneumonia is about 8- 9%. From an objective perspective, making general statements about the chance of death among young Malawi children with pneumonia would be speculation, albeit based on evidence: but all that can be said for certain is that if one of those particular 6,202 children were selected at random, the chance that child died was 8.4%.

The relationship between frequency data and objective probabilities will be further explored later.

The subjective interpretation

Bruno de Finetti, one of the most in fluential thinkers in the field, wrote

PROBABILITY DOES NOT EXIST

As Professor of the Theory of Probability, he was not dismissing his subject as a mirage, rather he rejected absolute claims such as 'The probability of Heads is one half'. To him, every statement involving a probability is just an expression of opinion, based on one's own experience and knowledge, and perhaps changing when more information arrives.

Consider the five assertions:

The England cricket captain will win the toss in England's next Test Match;

Whoever wins the Oscar for best actor next year will also win it the year after;

No person born in Oslo has yet won an Olympic fencing gold medal;

Richard III was responsible for the death of the Princes in the Tower;

Al Gore would have been elected US President in 2000 if Ralph Nader had not stood as a candidate.

To each of them, we can offer our degree of belief, or personal probability, or subjective probability. This will be some nonnegative number, not greater than 1: equivalently, it is a percentage between 0% and 100%, inclusive.

Zero and one represent, respectively, the two extremes of impossible , and certain . I am certain that the soccer World Cup will be hosted by an African nation again during the present century. I think it is impossible for someone under twenty years of age to win a Nobel Prize.

Assessing subjective probabilities

The five assertions above have different natures, and we have different kinds of evidence about them. For the first, we might appeal to symmetry between Heads and Tails. For the second, we have the history of the Oscars since 1929 to guide us. In both these cases, the truth or otherwise of the statement will become known within a finite time. The third is either true or false, and could be established now by a thorough trawl of Olympic records. The fourth is also either true or false, but we will never know which. We cannot rerun history to ascertain the truth or otherwise of the fifth claim.

Specific examples later will illustrate how subjective probabilities have been assessed. Aside from those arguments, there are at least three distinct general approaches. One is as the fair price for a bet that the event will occur. But this does not work for everybody: some people have principled objections to betting, others are unwilling to contemplate actions that might ever lead to a loss. And even for those who do feel comfortable with betting, their fair price might differ according as to which side of the bet they were on.

A second way to assess your degree of belief in an event uses the objective approach. Which offer would you prefer: to receive £5 if the event occurs, or to receive £5 if you correctly guess the colour, Red or Black, of the top card in a well-shuffled deck? If you prefer the latter, your degree of belief in the event is below 50%.

Suppose that is the case. Now compare the prospect of receiving £5 if the event occurs, or getting it if you correctly guess the suit of a randomly drawn card. The latter should occur 25% of the time, so your preference here will tell you whether your degree of belief is below 25%, or is between 25% and 50%.

More elaborate comparisons along these lines let you home in on a situation where you cannot say on which side your preference lies. Your degree of belief in the event will then be close to the objective probability of the corresponding card selection. Rather than use a deck of 52 cards, with its awkward fractions, you might think of an urn containing 20, or maybe 100, identical balls with which to specify the alternative events.

Give your answers with appropriate precision. Tennis players John Isner and Nicolas Mahut played the longest match in Wimbledon history in 2010; via counting, the chance that they would be drawn together again the next year (it happened!) is precisely 2 in 285, perhaps better rounded to 'a little under 1%'. But it was absurd of *Star Trek*'s Mr

Spock to tell Kirk that the odds against their escape in one episode were ‘approximately 7,824.7 to 1’.

For a third method, think of a modest sum of money, not so small that you are totally indifferent to it (say, one penny), nor so large that possessing it would make a dramatic difference to your circumstances (£1 million to most people, rather bigger for Bill Gates). For me, £10 fits the bill - call this unit amount. Now suppose that, somehow or other, the truth or falsity of the event will be revealed tomorrow: and you will receive this unit amount if it is true, but zero if it is false. But rather than wait for tomorrow, you could receive a definite proportion p of this unit amount today. (Getting the money today or tomorrow makes no difference to you.)

If p is tiny, you are likely to reject the offer, and will prefer to wait; if it is close to unity, you may well accept that definite amount. But there will be some intermediate value of p where you are indifferent between taking this offer, and waiting for the outcome to be revealed. This p is your degree of belief about this statement or event in question.

I offer my own subjective answers for the five assertions above. I can think of no sensible reason why one side should be more or less likely to win a cricket toss than the other, so my first figure is 50%. Looking at Oscar history, not only for actors but also the other categories, the award has

occasionally been repeated in successive years: perhaps there are more candidates these days, leading me to suggest 3%, or lower. Norwegians are not noted for fencing, but we have épée, foil, sabre, and the sport has appeared in all the Summer Games from 1896. Some native of Oslo might have won sometime, but I strongly doubt it - my figure here is about 95%. Prejudice in favour of the White Rose county, rather than objective evidence, leads me to suggest 10% for the fourth claim. For the fifth claim, considering the votes in each State, and thinking of a plausible division of the votes Nader received, guide me towards 20%.

Pause a while, and make your own suggestions for these five claims. The better you are at assessing probabilities when matters are uncertain, the more likely you are to be happy with the decisions you make in life.

Odds

Whether we use the classical approach, or frequencies, or degrees of belief, the term odds is often used when describing probabilities.

We might say that the odds of obtaining a Six with a fair die are ‘five to one against’ - for every time we get a Six in a sequence of throws, we expect not to do so five times. If an outcome is expected to be more likely than not, such as the higher ranked player winning a tennis match, that event is said to be odds on .

There is an exact correspondence between probabilities and odds, and we can switch easily between them. Thinking of frequencies can help. If the probability is 20%, or one fifth, we expect the event to occur in one occasion out of five, so the odds are 'four to one against'. For a probability of 75%, we expect it to occur three times out of four, giving odds of 'three to one on'. And if the odds are stated as six to five against, this indicates that for each five times the event happens, it fails to do so six times, so its probability is 5/11.

You do not have to stick to whole numbers. The probability that the top card in a well-shuffled deck is either a King or a Queen will be taken as 2/13. This could be quoted as 'eleven to two against' or, equally accurately, 'Five point five to one against'. Use whichever you like.

Although the phrase 'the odds are one to one' is never used, it would make perfect sense. It indicates that an event is expected to happen just as often as not, so its probability is one half. Instead, with a straight face, we say 'the odds are evens'.

Issues to resolve

There are no important disagreements about how to work with probabilities, but adherents of the three approaches we have described may deduce their values in different ways. Each perspective has its uses. In seeking to understand how

the subject works, we will appeal to whichever viewpoint appears appropriate.

The objective approach is limited to circumstances having finitely many outcomes, all judged equally likely. But no coin or die is perfectly symmetrical, and on what basis can we dismiss its imperfections as irrelevant? Can we even be sure that we agree on the number of possible outcomes? For example, suppose we are told that an urn contains two balls, either both White, both Black, or one of each colour. Should we argue we have three equally likely cases, or that there are really four equally likely cases, arising when the balls were inserted, in order, either as WW, WB, BW or BB? These different outlooks would give different answers for the chance that both balls are Black. Or suppose you reach a road junction with three possible exits, two of them leading to Newtown, the third to Seaport: making a 'random choice', is the chance that you aim for Seaport one third (one exit in three), or one half (one of two destinations)?

A frequentist seeks to deal with circumstances that are repeatable indefinitely often under identical conditions. The number of outcomes need not be finite - think of tossing the same coin until Heads appear three times in succession, or selecting a random point on a stick. But, however much care we take, the experimental conditions cannot be absolutely identical, and any limiting value can only be estimated. How should the error in this estimate be described? Claiming that the probability is at least 99% that the error is under 2%

requires a circular argument - we need to know what probability is, in order to define it!

For questions such as the probability that one country invades another, or the chance that a particular heart transplant is successful, the circumstances arise once only, and the alternatives cannot be reduced to a finite list of equally likely cases. The objective and frequency approaches are silent on these matters. A subjective approach is required.

A subjectivist must ensure that her beliefs are consistent with each other. For example, in the UK National Lottery, a machine selects six numbers from the list $\{1, 2, 3, \cdots, 49\}$, and Susie may be content to take all 14 million or so possible selections as equally likely. Then, when asked which is more likely,

(a) that no number drawn exceeds 44, or

(b) that those drawn do not include two consecutive numbers,

she may, after a little thought, come down on one side or the other. But if she selects either of these events as more likely than the other, she will be guilty of inconsistency, as proper counting shows that they can occur in exactly the same number of ways! Nothing in the subjective approach specifies how such an inconsistency should be resolved, merely that it must be.

Because we wish to think about probabilities in circumstances wider than when there are finitely many equally likely choices, and in circumstances that cannot be repeated indefinitely often, we will take the subjective approach as the default option. But we are likely to hold more firmly to our opinions when they are backed up by either an objective, or by a frequency, argument.

Interpretations

Using the 'balls in a bag' viewpoint, the probability of some event is taken as the proportion of Red balls in the bag. So a value of zero can occur only if there are no Red balls, in which case the event will never happen. Similarly, a probability of unity corresponds to every ball being Red, so here the event occurs every time. These values, zero and unity, are the only ones that can be conclusively proved wrong by experimental evidence: if the event happens, its probability cannot be zero, if it fails to happen, its probability cannot be unity. And this is true for the frequency, or subjective approaches also. So suppose the probability has some intermediate value, say 3/4.

We first dispose of one finicky point. No matter how well a roulette wheel has been engineered, it is physically impossible that all the numbered slots have exactly the same chance. What the casino requires is that the chances are close enough to the ideal that it is inconceivable that any number could be picked out as more or less likely than

another. Similar remarks apply to dice, coins, or cards. So statements like 'the probability is 3/4' will mean that the probability is close enough to 3/4 for all practical purposes. Otherwise, a pedant might smugly tell you that he knows that the probability is not 3/4, without fear of contradiction.

In the context of repeatable experiments, what do we expect to follow from the claim: 'The probability of a Red ball is 3/4'? Emphatically, we do not expect that if we conduct this experiment four times (replacing the ball drawn on each occasion), we shall draw a Red ball in precisely three of them. It is possible that four repetitions throw up no Reds at all, or even that Red happens every time. But over a long series of repetitions, we do expect the overall frequency of Red to be close to 3/4.

There are no black/white answers to what constitutes a long series of experiments, nor to how close to 3/4 is acceptable. If I obtained Red only 20 times in the first 40 repetitions, I would have very strong doubts about a claim that the probability was 3/4; but those doubts would be largely assuaged if the next 40 repetitions gave 28 Reds. Believing or disbelieving this claim can be a provisional position for quite some time. Assuming the experimental conditions do remain unchanged throughout, use all the data collected to reach a decision - short runs can mislead.

I will offer some guidelines, and justify them later. Take the case when we make one hundred repetitions, and the supposed probability is some middling value, near one half. Compute the difference between this figure and the actual frequency from data: if this difference exceeds 0.1, I would have some doubts about the claim, and if it exceeded 0.15, I would have strong doubts. With a thousand repetitions rather than a hundred, I expect closer agreement, so replace those numbers by 0.03 and 0.05. If the supposed probability is closer to zero or unity, say 10% or 90%, I would also require better agreement. It can be much easier to be convinced, on the basis of repeated experiments, that a particular probability is not some alleged value.

What about a subjective assessment, such as that the probability of rain tomorrow is 60%? We cannot recreate today's weather conditions hundreds of times, and check how often it rains. This 'experiment' can be conducted once only. But we might test the claim by looking at the process that led to it being made. Forecasters use models of weather patterns to reach their conclusions, and even if the figure on their computer screen is 31.067%, they sensibly offer round figures. You hear 'The chance of rain is about 30%'. So now you can collect data for different days, and look at the empirical evidence - in how many of the 83 days last year when the chance of rain was put at 30% did it actually rain? So long as that proportion was reasonably close to 30%, your belief in the method is reinforced, so accepting the figure given for 'tomorrow' is a rational response.

Probability is the key to making decisions under conditions of uncertainty. If you honestly believe that the probability of a particular event or statement is unity, you should act as though it will definitely occur; and if your honest belief is that the probability is zero, act as though it cannot occur.

If you think the probability is some value between zero and one, act as though you expect it to occur that proportion of the time.For example, if your judgement is that the probability is 60%, imagine that you will face this situation a hundred times, in sixty of which (but you have no idea which sixty) this event will happen, and forty times it will not. Swallow hard, and decide on your action, taking into account this balance. Had you judged the probability to be 80%, so that you expect the event to happen rather more often, your action might well be different.

As Bishop Joseph Butler wrote in his 1736 Analogy of Religion , ‘To us, probability is the very guide to life’ .

Chapter 2 The workings of probability

As well as the subjective, objective, and frequentist approaches to probability, there are other standpoints. For example, should one always insist on associating a probability with a number? Might it be enough to say that one probability was greater, or one degree of belief was more intense, than another? And should we necessarily offer an initial set of axioms – self-evident truths – on which to erect a theory?

Many distinguished writers have felt it useful to have two separate approaches, one for degrees of belief and one for objective probabilities. Both would have the same rules of logic, free from contradictions, but how values of probabilities were arrived at, and how they are interpreted, could differ. Any theory should be consistent with the classical view, based on repeatable experiments with equally likely outcomes. So we will focus on that case, seeking any rules that the notion of probability must obey.

The Addition Law

Deal one card from a well-shuffled pack. We take all cards as equally likely, so the probability of any event, such as obtaining a Club, or a Spade, or an Ace is found by calculating the proportion of all possible outcomes that lead to those events. How might we find the probability that either of two such events occur?If those events have no

outcomes in common, we say that they are mutually exclusive , or disjoint . The events ‘Get a Spade’ and ‘Get a Club’ are disjoint, but the events ‘Get a Spade’ and ‘Get an Ace’ are not disjoint, as the Ace of Spades belongs to both. When two events are mutually exclusive, then the total number of outcomes that lead to either event is just the sum of the numbers for each event separately, so we have a simple result: whenever two events are mutually exclusive ,

the probability that at least one occurs is the sum of their individual probabilities.

This is the Addition Law . It plainly holds in all experiments where we would take the classical view: using the balls in a bag analogy, it is the same as saying that the total number of balls that are either Red or Blue is the sum of the number of Red balls and the number of Blue balls. And in any repeatable experiment, such as rolling dice,

or spinning roulette wheels, the sum of the individual frequencies of two disjoint events is inevitably the frequency that at least one of them occurs. So frequentists accept the Addition Law too.

Also a subjectivist accepts this Law. For otherwise, there would be two disjoint events, call them A and B, where it did not hold. In that case, the subjectivist could be confronted by three bets: one about A, one about B, and one about either A or B, and would accept each bet on its own as fair. But he could be guaranteed to lose money if all three

bets were struck! The Addition Law forbids this inconsistency.

This Addition Law extends to a collection of many events, provided no two of them have any outcomes in common - they are pairwise disjoint . The probability that at least one among even millions of pairwise disjoint events occurs is just the sum of their individual probabilities. But suppose the number of outcomes is no longer finite: for example, tossing an ordinary coin repeatedly until Heads appears for the first time.

The possible outcomes of this experiment are the unending list {1, 2, 3, 4,….}, each with its own non-zero probability. What is the chance that we take some even number of throws to get a Head? That event happens when any of the outcomes {2, 4, 6, 8, ...} happen. Could we compute its probability by adding up the corresponding individual probabilities?

There is no mathematical difficulty in doing this adding up, but that action falls outside the scope of the classical view of probability, which deals only with a finite list of outcomes. There is no consensus as to whether the Addition Law for such an unending list should be part of the workings of probability. In favour of including it, we may be able to find the probabilities of a wider class of events than without it; against inclusion, as it is not part of the classical theory, we should be cautious about taking steps

that might have hidden pitfalls. There is no right or wrong answer.

I’m a pragmatist. I am content to extend the use of the Addition Law in this way, and I have never felt uncomfortable with the results of doing so. This position is a standard part of the account given in most books used to teach the subject at university. But de Finetti took the cautious view to avoid making this extension, and others have felt the same way.

The Multiplication Law

If you toss an ordinary coin, you will expect to guess Heads or Tails correctly half the time. If you shuffle a deck of cards, and predict whether the top card is Red or Black, you also expect to be correct half the time. When you guess both a coin toss and a card colour, how likely are you to get both correct?

Think of conducting this double experiment a hundred times. You expect to guess the coin correctly about fifty times, and when you do so, you expect to go on to guess the card colour half the time.

That suggests you expect to get both right on about twenty-five occasions, and it looks sensible to offer 25%, or 1/4, as the chance of being right both times. For these experiments, the chance of being correct both times is found just by multiplying their individual chances.

Ten balls of the same size and composition are labelled with the numbers zero up to nine, and one of them is selected completely at random. So it is equally likely to show a Low number (zero to four), or a High number (five to nine). Five of these numbers are coloured Green, the rest are Blue, so Green and Blue are also equally likely. Trying to guess the colour, or whether it is Low or High, we have a 50% chance either time. What about the chance that the ball we draw is both Low and Green?

The argument above with the coin and the cards suggests 1/4 as the answer, but a moment’s thought shows this cannot be correct. With ten balls, it is impossible that 1/4 of them (two and a half!) will be both Low and Green! The correct answer depends on which numbers are coloured Green, and which Blue. So suppose numbers one to five are Green, the rest are Blue.

In that case, four of the ten numbers (one, two, three, and four) are both Low and Green, so the chance is 0.4. But, as we did with the first problem, we can also use a two-stage process: in one hundred repetitions of this experiment, we expect to get a Low number fifty times. Four of the five Low numbers are Green, so when we did get a Low number, we expect it to be Green 4/5 of the time. Overall, we expect a Low Green number forty times, leading again to the answer 0.4.

With the coin and cards, the outcome of the coin toss has no bearing on the card drawn. We do not change our minds

about the chance of a Red card when told whether the coin falls Heads - the conditional probability of the second event, given the first, is just its ordinary probability. When this happens, the two events are said to be *independent* , and the chance that both occur is the product of their individual probabilities.

With the ten balls, the chance both events occur also arises as a product, in which the first component is also the probability of one event (Low number), and the second is the conditional probability of Green when this happens. So really the two calculations are identical in form, the only difference is whether the outcome of the first event affects the chance of the second. Each time, we have used the *Multiplication Law* , which says:

the probability that both of two events occur is the probability of the first, multiplied by the probability of the second, conditional on the first happening.

Independence

We used the term 'independent' to describe the circumstances when the occurrence of the first event does not change our assessment of the chance of the second. Suppose this holds, but we learn that the second event has happened; might this affect our assessment of the chance of the first?

No. Whenever the fact that one event has or has not occurred makes no difference to the chances of another event,

it turns out that whether or not this second event occurs makes no difference to the chances of the first. Two events are independent when the occurrence or non-occurrence of either makes no difference to the probability of the other. To find the chance both occur, multiply their individual chances.

Events that have no bearing at all on each other, like rain today in Tunis and the gender of the next birth in Paris, are surely independent. But sometimes independence is not obvious. Using an ordinary fair die, consider the events ‘Get an even number’ and ‘Get a multiple of three’, with respective chances one-half and one-third. The only way both occur is when we get a Six, having probability one-sixth. And since multiplying one-half and one-third gives one-sixth, those two events are independent. The chance of getting an even number does not change if we are told whether or not a multiple of three occurs (and vice versa).

Now consider the same problem when you have an eight-sided fair die, or a ten-sided fair die, with the sides labelled one to eight or one to ten respectively. Do the arithmetic: you should find that the two events are independent in one of these cases, but not independent in the other. Intuition about independence is useful, but not always enough.

Assuming two factors to be independent, when they are not, is one of the most common mistakes made in assessing

probabilities. Suppose that, in a university's graduate school, half the students are female, and one in five study engineering. Select one student at random: the probability this student is female can be taken as one-half, the probability the student studies engineering will be one-fifth. However, you will find that the probability the student is a female engineer is much less than their product, one-tenth.

Events with overlap

The Addition Law shows how to find the chance of at least one of two events, provided they are disjoint. What if they are not disjoint? For example, in drawing one card at random, what is the chance it is either a Spade or an Ace? The Ace of Spades falls in both categories, so if we just added the respective probabilities, we would count that card twice. To correct for the outcomes that would be double-counted and find the probability that at least one of two events occurs,

add their individual chances, then subtract the chance both occur. Of course, if the two events are disjoint, it is impossible that both happen, so this extra term has value zero, and we are back to the original Addition Law.

Let's see this notion in action in the two earlier examples. With the coin and Red/Black card, the chance we get at least one guess right comes from the arithmetic 1/2+1/2-1/4, which is 3/4. In the other example, the chance a

randomly chosen numbered ball is either Low or Green is 1/2+1/2−0.4 = 0.6.

And the chance of drawing either a Spade or an Ace arises as 13/52+4/52−1/52 = 16/52, confirmed by noting that exactly 16 of the 52 cards satisfy this condition.

This last calculation should warn you against making early arithmetic simplifications. Yes, 13/52 is the same as 1/4, and 4/52 is the same as 1/13, but to add 1/4 and 1/13, you are better off with their original fractions. And it is seldom helpful to re-write a friendly fraction like 5/13 as its ugly decimal approximation 0.38461538 . . .

More than two events

A collection of many events is described as independent whenever knowing whether or not some of them occur makes no difference to the probabilities of any of the others. In this case, the Multiplication Law means that, whatever selection of events we make from this collection, the probability that all of them occur is just the product of their individual probabilities.

But how might we find the probability that three or more events all occur, when they are not independent? For example, in whist or bridge, the cards are randomly shuffled and shared equally among four players. How likely is it that they all receive exactly one Ace?

Consider the four separate events: Anne gets exactly one Ace; Brian gets one Ace; Colin has one Ace; Debby gets one Ace. Plainly, these four events are not independent, as if any three of them happen, the other is certain. We will find the probability they all occur via a three-stage process.

First, we find the probability that Anne has exactly one Ace. Assuming all possible ways of dealing the cards are equally likely, we have an exercise in counting: count the total number of possible deals, and then count in how many of them she gets exactly one Ace. Believe me, the chance works out as just under 44%.

Assume Anne has just one Ace (and hence twelve non-Aces). That leaves three Aces and thirty-six non-Aces for the other players, and Brian gets thirteen of them, chosen at random. A similar counting exercise on this smaller pack shows that the chance he would get exactly one Ace is just over 46%. The Multiplication Law then tells us that the chance of both events, i.e. that both Anne and Brian have exactly one Ace, is the product of these two values, just over 20%.

So now assume Anne and Brian each have exactly one Ace. Then Colin receives thirteen cards at random from the two Aces and twenty-four non-Aces that remain: the chance that he gets exactly one Ace is found to be 52%.

The final step is to use the Multiplication Law once more, to combine these last two calculations: the chance that Anne and Brian, and then also Colin, all have exactly one Ace

is a little over 10%. If this happens, Debby inevitably has the final Ace, so we have found the answer we seek.

This answer itself is of no real consequence – although the deal is totally random, the most equitable outcome for the Aces is rather unlikely – but the method used is universal. To find the chance that every event in a collection occurs, break things down into stages. Find the chance for one event; then, assuming this event occurs, find the chance of a second; now assuming both of these occur, find the chance of a third; then assuming all these three events happen, find the chance of a fourth – and so on. Finally, multiply all these quantities together.

Where else might we have to follow this path? Suppose my journey has three stages, and I can assess their separate chances of having no delay: however, all the stages will be affected by the weather, and delay on one stage will change the chance of delay elsewhere. In manufacturing industry, the safety of a piece of equipment will rely on several components which do not operate independently – some may use the same water supply, others may have been inadequately tested by the same unreliable employee. With a medical procedure, whether or not the things that can go wrong are independent of each other can make a huge difference to the overall chance that all turns out well.

If events are independent, then the chance they all occur is just the product of their individual chances. But we are

seldom lucky enough for this condition to hold: a stage-by-stage assessment, with probabilities changing as the work progresses, is the norm.

What about the chance that at least one of three or more events occurs? The Addition Law does extend to this case, but as the expression is cumbersome, I will not write it down. Its recipe follows the same path as described when using the Multiplication Law for the chance that all of many events occur: take it one step at a time.

Using the word independent when disjoint is meant, and vice versa, are common errors. The example of choosing one card at random can help you see how to avoid it. Here, the events 'Get a Spade' and 'Get a Club' are disjoint, but far from independent, as if either occurs, the other cannot, so the chance both occur is zero! Also 'Get a Spade' and 'Get an Ace' are independent (yes?), but plainly not disjoint.

Remember: the Addition Law is used to find the chance of at least one event, the Multiplication Law is used to give the chance they all occur.

It is sometimes said that counting really goes one, two, infinity. This aphorism carries the truth that if we can make the step from dealing with one case to dealing with two cases, then subsequent steps to three, four, five, etc. cases are trivial in comparison. This surely applies to both the Addition and Multiplication Laws.

A neat trick

Any event either happens, or it does not. The total probability is split between the event happening, and it failing to happen. So if we can find the chance an event does not happen, we can deduce the chance it does occur by subtraction from 100%.

To illustrate, let’s find the chance of at least one Six when an ordinary fair die is rolled twice. Any outcome is a pair of numbers showing the scores of the first roll, then the second, e.g. (5, 2) or (4, 4), and we take all such outcomes as equally likely. Each roll has six possible results, leading to 6x6=36 outcomes altogether. Our event does not happen when neither die shows a Six, for which there are 5x5=25 outcomes. The chance of no Sixes is 25/36, so the chance of at least one Six is 11/36, a bit less than one in three.

This leads on to a junior version of a gambling problem solved by Blaise Pascal and Pierre de Fermat in 1654. How often must we roll a die to make it more likely than not that we get at least one Six, i.e. that the chance of getting a Six is more than one half? We have just seen that two rolls are not enough.

Each extra roll increases the number of possible outcomes by a factor of six, while the number of outcomes without a Six gets multiplied by five. So a third throw generates 216 outcomes in all, and 125 of them - over half - contain no

Six; three rolls are not enough either. However, four rolls give 1,296 outcomes, and only 625 of them have no Sixes - fewer than half. That leaves more outcomes that include a Six than outcomes with no Sixes, so now a Six is more likely than not. Four rolls suffice.

The actual game analysed by Pascal and Fermat involved rolling not one die, but two dice together; and asking how often this needs to be done to make it more likely than not that a Double-Six turns up at least once. The method of solution is the same, but the raw arithmetic is formidable. Today, we can quickly reach the answer with a microcomputer or a pocket calculator, while logarithms and slide rules had conveniently just become available in the 17th century. With up to 24 rolls, it is more likely than not that no Double-Six appears, but a 25th roll tips the balance the other way.

Most problems of the form 'Find the chance of at least one of these events happening' are best solved in this manner: work out the chance that none of them arises, and then subtract from unity.

Chapter 3 Historical sketch

Beginnings

A game popular in Florence around 1600 rested on the total score from three ordinary dice. The scores of Three, when all dice scored one, and Eighteen when they all scored six, arose rarely, with most scores near the middle of the range. You should check that there are six different ways of scoring Nine (e.g. 6+2+1, 5+2+2, etc.), and also six ways of scoring Ten. It was commonly believed that this 'ought' to make totals of Nine or Ten equally frequent, but players noticed that, over a period of time, the total of Ten occurred appreciably more often than Nine. They asked Galileo for an explanation.

Galileo pointed out that their method of counting was flawed. Colour the dice as Red, Green, and Blue, and list the outcomes in that order. To score Nine from 3+3+3 requires all three dice to show the same value, and that can happen in one way only, (3,3,3). But the 5+2+2 combination could arise as (5,2,2), (2,5,2),

or (2,2,5), so this combination will tend to arise three times as often as the former; and 6+2+1 arises via (6,2,1), (6,1,2), (2,6,1), (2,1,6), (1,6,2), and (1,2,6), so this combination has six ways to occur. A valid approach to how often we can get the different totals takes this factor into account, and does indeed lead to more ways of obtaining Ten

than Nine. The Florentine gamblers learned a vital lesson in probability - you must learn to count properly .

In the summer of 1654, Pascal (in Paris) and Fermat (in Toulouse) had an exchange of letters on the problem of points . Suppose Smith and Jones agree to play a series of contests, the victor being the first to win three games; unfortunately, fate intervenes, and the contest must end when Smith leads Jones by 2-1. How should the prize be split?

Such questions had been aired for at least 150 years without a satisfactory answer, but Pascal and Fermat independently found a recipe that, for any target score, and any score when the contest was abandoned, would divide the prize fairly between them. They took different approaches, but reached the same conclusion, and each showered praise on the other for his brilliance. For the specific problem stated, the split should be in the ratio 3:1, with Smith getting 3/4 of the prize, Jones 1/4.

The essence of their solution was to suppose that both players were equally likely to win any future game. They counted how many of the possible outcomes of these hypothetical games would give overall victory to either player, and proposed dividing the prize in the ratio of these two numbers. In different language, the prize should be split as the ratio of the two probabilities of either player winning the series, assuming they were evenly matched in future games. The systematic study of probability had begun.

This issue was settled via the objective approach to probability, but Pascal also thought more widely. He suggested a wager about the existence of God. 'God is, or is not. Reason cannot answer. A game is on at the other end of an infinite distance, and Heads or Tails is going to turn up. Which way will you bet?'

He argued that if God exists, the difference between belief and unbelief is that between attaining infinite happiness in heaven, or eternal damnation in Hell. If God does not exist, belief or unbelief lead to only minor differences in earthly experience. Thus an agnostic should lean strongly to belief in God.

In this game, the values of the chances of 'Heads' or 'Tails' are personal choices, not derivable from symmetry or counting arguments. Thus Pascal was a pioneer in the subjective approach to probability too.

The Swiss Family Bernoulli

During the 17th and 18th centuries, members of the Bernoulli family from Basle made significant advances in mathematics, including probability. Rivalry was a spur: one of them would pose challenges, another would respond, the originator of the challenge would claim to find flaws in the supposed solution, and so on.

Games of chance inspired much of the early interest in the workings of probability. In these games, be it rolling

dice, dealing cards, or tossing coins, some 'experiment' is carried out repeatedly under essentially the same conditions. The natural question, raised earlier, is: how does the observed frequency of an outcome relate to its objective probability?

Jacob Bernoulli gave an answer in his posthumously published The Art of Conjecturing (1713), nicely illustrated by his example. Suppose 60% of the balls in an urn are White, the rest are Black, and one ball is drawn at random. That ball is replaced, and the experiment repeated many times. Bernoulli showed that, so long as at least 25,550 drawings are made, for every time the proportion of White balls falls outside the range from 58% to 62%, it will fall inside that range at least one thousand times. Informally, the observed frequency of White balls is, in the long run, overwhelmingly likely to be close to its objective probability.

A similar analysis applies to any experiment that can be repeated indefinitely under identical conditions, where the result of one experiment has no effect on the others. Each time, certain outcomes denote Success, and their objective probability is some fixed value p . (This notion now carries the label Bernoulli trials .) Take any interval, as small as you like, around the value p - plus or minus 2%, plus or minus 0.1%, it matters not. Also, say how much more often you want the running frequency of Successes to be inside this interval, rather than outside it - a hundred times as often,

a million times, whatever. Bernoulli’s methods show that any such demand can always be met, provided the experiment is repeated often enough. The observed frequency will be as close to

the objective probability as you like, given enough data. This assertion is known as the Law of Large Numbers . The family’s fame was honoured in 1975 by the name choice ‘The Bernoulli Society’ for an international society whose main purpose is to foster advances in the study of probability and mathematical statistics.

Abraham de Moivre

De Moivre settled in England as a Huguenot refugee, and made a living from chess and from his knowledge of probability. Isaac Newton, then over 50 years old and with many calls on his time, deflected enquiries about mathematics with the words ‘Go to Mr de Moivre, he knows these things better than I do.’ De Moivre’s Doctrine of Chances appeared in English in 1718, and its second edition, in 1738, contained a major advance on Bernoulli’s work. To appreciate what he did, consider something specific: if a fair die is rolled 1,000 times, how far from the average frequency can we reasonably expect the number of Sixes to be?

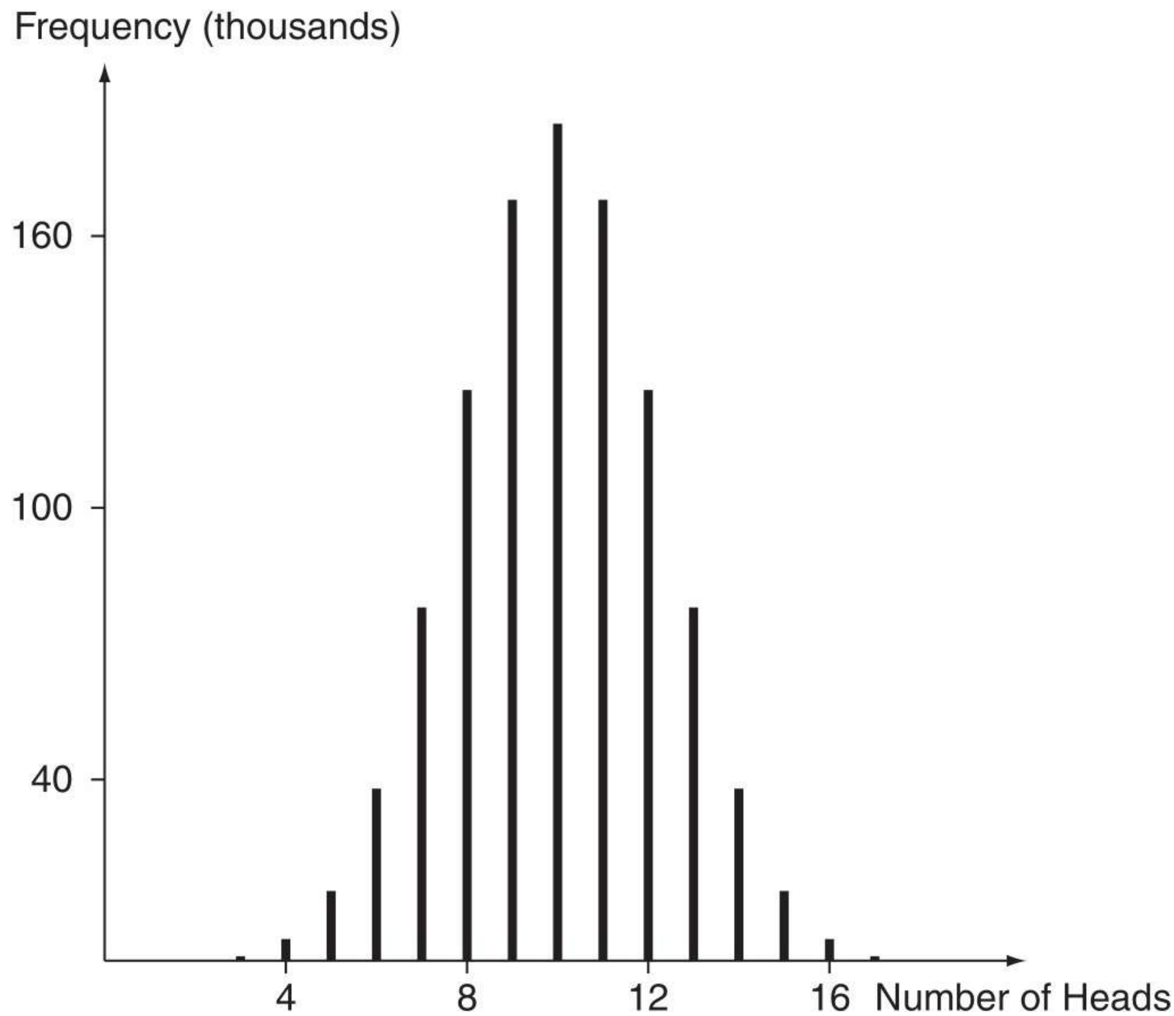
De Moivre developed a simple formula that was widely useful for questions of this nature. One of his superb insights was to realize that the deviation of the actual

number of Sixes from the average expected was best described by comparing it to the square root of the number of rolls.

It is hard to overplay the significance of this discovery. When you hear that an opinion poll has put support for a political party at 40%, it is often accompanied by a reminder that this is only an estimate, but that the true value is 'very likely' to be in some range like 38% to 42%. The width of such a range tells you about the precision of the initial figure of 40%, and if you want higher precision, you need a larger sample: this square root factor means that to double the precision, the sample needs to be four times as large! We have a law of diminishing returns with a vengeance - to do twice as well, we must spend four times as much.

De Moivre's approach can be illustrated by looking at how many Heads will occur in twenty throws of a fair coin. Taking all sequences of length 20 such as HHHTH . . . HTHT as equally likely, we can construct Figure 1 , where the heights of the vertical bars show how many of the one million or so different sequences produce exactly 0, 1, 2, . . . , 19, 20 Heads. The respective objective probabilities are then proportional to these heights. De Moivre showed that the best-fitting continuous smooth curve through the tops of these bars is very close to a particular form, now often called the normal distribution .

A curve of this nature arises for any large number of coin throws, and also when the chance of Heads differs from one half. All these curves bear a simple relation to each other, so de Moivre could produce a single numerical table for just one basic curve, and use it everywhere. A good estimate of the proportion of times that the overall frequency of Successes would be within certain limits could now easily be found - all that was needed was the chance of Success, and the number of times the experiment was to be conducted. You're going to roll a fair die 200 times and you want to know how likely it is that the number of Sixes will be between



1. Relative frequencies of Heads in 20 throws

30 and 40? Or how likely is it that a fair coin will fall Heads more than 60 times in 100 tosses? No problem – de Moivre had the solution.

Suppose we know the ages of death for a group of men, all of whom reached at least their fiftieth birthday. De Moivre's work could answer the question: 'If a man aged 50 is more likely than not to die before reaching 70, how likely is it that the figures observed for that group would arise?'

Useful though this was, it did not answer the key question posed by the nascent life insurance industry: ‘How sure can we be that a 50-year-old man is more likely than not to die before he reaches the age of 70?’

Inverse probability

The ideas of Thomas Bayes, a Presbyterian minister who dabbled in mathematics, are far better appreciated now than in his lifetime. His Essay towards solving a problem in the doctrine of chances , published in 1764, three years after he died, gives the beginnings of a general approach to subjective probability, and a way of addressing the actuaries’ problem about inferring probabilities from data. It also included an essential tool for working with probabilities, termed Bayes’ Rule.

To illustrate the latter, suppose we throw a fair die twice. Given that the score on the first throw is three, it is easy to find the chance that the total score is eight, as this happens precisely when five is scored on the second throw. With hardly a pause, we give the answer 1/6. But turn the problem round, and ask: given that the total score is eight, what is the chance that the first throw yielded three? The answer is far less obvious, but can be found by applying Bayes’ Rule. Under the standard model of dice throws, the chance turns out to be 1/5.

This notion of inverse probability is central to the way evidence should be considered in criminal trials. Suppose

fingerprints found at a crime scene are identified as belonging to a known individual, Smith. The probability of finding this evidence, if Smith is innocent, is likely to be very low. But it is not ‘How likely is this evidence, given that Smith is innocent?’ that the Court passes judgment on: it is ‘How likely is Smith to be innocent, given this evidence?’ Bayes’ Rule is the only sound way to obtain an answer. We will see in later chapters how this Rule helps in making sensible decisions.

The insights shown by Bayes were overlooked for many years, but he did identify the central problem: if the chance of Success in a series of Bernoulli trials, like dice throws, is unknown, but the respective numbers of trials and Successes are known, how likely is it that this unknown chance falls between specified limits? Laplace, a far superior mathematician, was able to carry out the computations that had defeated Bayes.

From tentative beginnings in 1774 to a synthesis in 1812, Laplace steadily improved his analysis, and gave explicit formulae to answer Bayes’ question. For example, using data on the numbers of male and female births in Paris, he concluded that it was beyond doubt that the chance of a male birth exceeded that for a female - he put the probability this was false as about 10 - 42 !

Bayes in buried in the London cemetery of Bunhill Fields, near the Royal Statistical Society. The vault has been

restored, and displays a tribute to Bayes paid for by statisticians worldwide.

The Central Limit Theorem

Write the list of outcomes of a collection of Bernoulli trials as a sequence of Successes and Failures, e.g. FFFSF FFSSF SFF . . . Now replace each S by the number one, and each F by zero, giving 00010 00110 100 . . . This indicates a cunning way to think about the total number of Successes in these trials: it is just the sum of these numbers (agreed?). De Moivre had given a good approximation that described how this sum would vary, using his so-called normal curve.

A vast array of quantities we might want to consider do arise as a sum of randomly varying individual values. For example, a local authority responsible for rubbish disposal is interested mainly in the total amount over the town, and not in the separate random amounts from each household. When a gardener sows runner beans, it is the total yield, not that in each pod, that concerns him. A casino judges its financial success on its overall winnings, irrespective of the fates of individual gamblers. Being able to regard an item of interest as the sum of a large number of random bits is often fruitful.

Laplace extended de Moivre’s work to cover cases like these. He established a Central Limit Theorem , which says that something that is the sum of a large number of random bits will, to a good approximation and in a wide range of

circumstances, fit de Moivre's normal distribution. We don't need the details of how the individual components tend to vary, the way the total amount varies will closely follow this normal law.

To use this idea, we require just two numbers: first, the overall average amount, and second a simple way of expressing its variability. Given those two numbers, any probability can be found from de Moivre's tables.

Enter Carl Friedrich Gauss (1777- 1855), bracketed with Newton and Archimedes at the top of the mathematical tree of genius. He was investigating how to deal with errors in the observations of the positions of the stars and planets. He suggested that on average the error was zero - observations were just as likely to be wrong a bit to the left as a bit to the right - and its size followed this same normal distribution. He took this path for its mathematical simplicity, but when Laplace saw Gauss's book, he linked it to his own work. He argued that because the total error in an observation arises as the agglomeration of many random factors, such an error ought to follow the normal law. Gauss's lame excuse of 'mathematical convenience' was replaced by Laplace's more persuasive 'mathematics indicates that . . . '.

The term 'normal', applied to this distribution, is unfortunate. It suggests that, in the first instance, we should expect any data we come across to follow its format,

but this is far from the case. To avoid this implication, and to honour a great man, we will switch to the alternative term Gaussian . If you can persuade yourself that your item of interest can plausibly be regarded as the sum of a large number of smaller variable bits, having largely unrelated origins, this Central Limit Theorem says that the item can be expected to vary in a Gaussian manner.

Do observational errors really follow this law? According to Henri Poincaré, the last mathematician to feel comfortable across the whole existing mathematical spectrum, 'Everybody believes in it, because the mathematicians imagine it is a fact of observation, and observers that it is a theory of mathematics.'

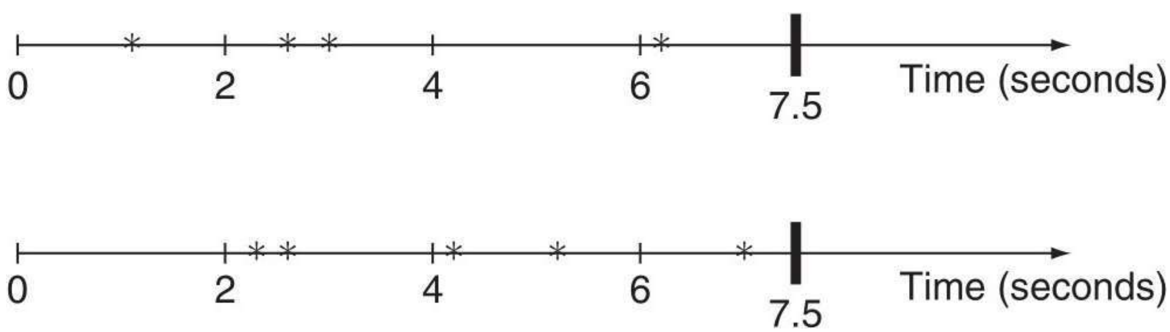
Siméon Denis Poisson

Poisson is best known for a distribution - the way in which probabilities vary around an average - that carries his name. An example arose in the work of the physicist Ernest Rutherford and his colleagues, when they counted how many alpha particles were emitted from a radioactive source over intervals of length 7.5 seconds. This number varied from zero to a dozen or so, with an average just under four. Figure 2 illustrates two typical experiments, showing (in those cases) four/five emissions. Rutherford expected the emissions to occur at random.

Chop the 7.5 seconds into a huge number of really tiny intervals, so small that we can neglect the possibility of

more than one emission in them. All but a few intervals will have zero emissions, the rest will have just one. Within each tiny interval, regard an emission as a Success, so that the total number of particles emitted is just the number of Successes – Bernoulli trials again.

In a really tiny interval, the chance of a Success is effectively proportional to its length, so as this length shrinks, we have an increasing number of intervals, each having a decreasing chance of Success. Poisson worked out the exact chances for 0, 1, 2, . . . emissions altogether as the lengths of the tiny intervals reduce down to zero.



2. Times of emission of alpha particles

This Poisson distribution arises frequently, at least as an excellent approximation, whenever the things we count are happening 'at random'. It was appropriate for Rutherford's data; it fits the numbers of flying bombs that landed on different parts of south London in World War II; it seems useful as a model for the number of misprints in each block of 1,000 words in a book. If you simultaneously deal out two randomly shuffled decks of cards, face up, on average you

will have exactly one match between them; but the actual number of matches will vary very much like a Poisson distribution. A gruesome example of this distribution, foisted on generations of students, is of twenty years of data for the numbers of officers in the different Prussian Cavalry Corps who were kicked to death by their horses.

All of those examples conform to the same pattern: a large number of opportunities, each with a tiny chance of coming off. Whenever the phenomenon you are studying fits that template, this Poisson model is likely to be useful.

The Russian School

A mathematical theorem takes the format: if certain assumptions are true, then a desired conclusion follows. The main interest is in applying the desired conclusion, so it is most useful when the required assumptions are not very onerous. Sometimes the desired conclusion can be demonstrated only under very restrictive assumptions, or with great difficulty: later workers may find easier ways to use the same assumptions, or reach the same conclusion under less restrictive conditions. Best of all is when the conclusion can be shown true under very mild assumptions, and with a short and elegant argument. The work of Pafnuty Chebychev (1821-94) gives a fine example of this ideal.

Chebychev helped to show how a Law of Large Numbers applied in wider circumstances. The original Law related to Bernoulli trials, describing how well the proportion of

Successes in a sequence of trials could estimate the chance of Success. If we want to estimate the average height of soldiers joining an army, or the cost of feeding a family for a week, it seems obvious that we can do so by taking a suitable sample from the relevant population. But how good will that estimate be? Chebychev’s work gave a firm idea of the probability that the error would be small enough to make the estimate reliable.

Much of Statistics rests on the applications of these ideas.

Chebychev’s best-known student is Andrey Markov, whose teaching inspired a further generation of talented Russians. Markov applied his ideas to poetry and literature. By replacing the vowels and consonants in Pushkin’s Eugene Onegin by the respective letters v, c, he generated a sequence with just those two symbols. In the original Cyrillic alphabet, vowels formed about 43% of the text. After a vowel, another vowel occurred some 13% of the time, while after a consonant, vowels arose 66% of the time.

To predict whether the next symbol would be v or c, he discovered that, given the current symbol, he could effectively ignore all its predecessors, so little help did they give.

This ‘forgetting’ property holds widely. Examples include: the successive values of a gambler’s fortune; the daily weather (Wet or Dry) in Tel Aviv; the lengths of many

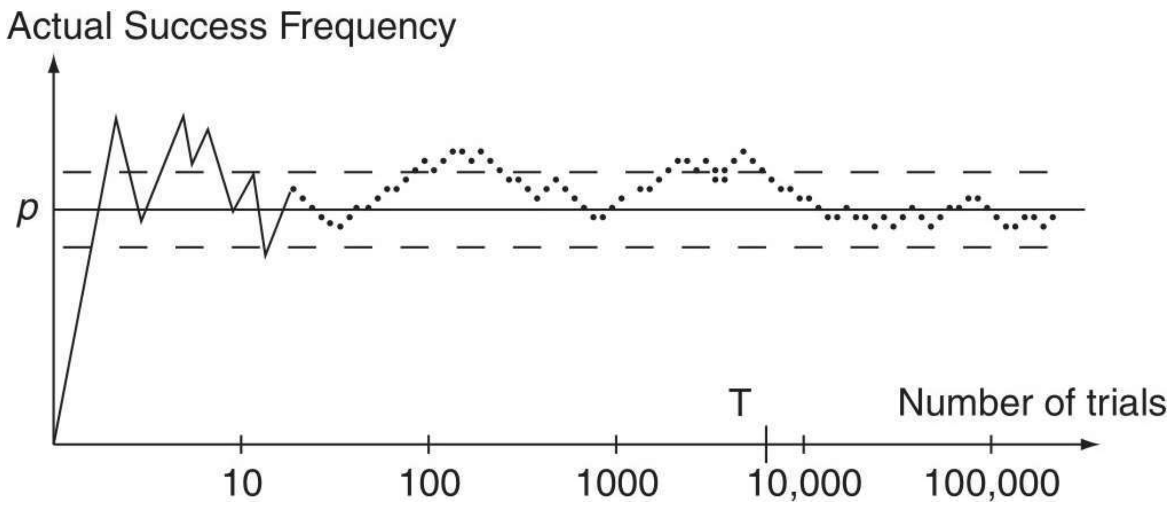
queues, counted as each customer leaves; the genetic compositions of consecutive generations; the diffusion of gases between two linked containers.

Whenever in a randomly varying sequence where we wish to predict future values, we can, knowing the present, ignore earlier values, the sequence is said to have the Markov property. The theory of such sequences is now well developed, and is the basis for many successful applications of probability ideas.

Markov, also active in politics, had a fine sense of mathematical history. In 1913 the Russian government organized celebrations to mark 300 years of Romanov rule, so Markov countered with events to commemorate the 200 years since Bernoulli's discovery of the first Law of Large Numbers.

I digress to mention the work, early in the 20th century, of the Frenchman émile Borel. Recall the Law of Large Numbers for Bernoulli trials: that after many trials, it is overwhelmingly likely that the actual frequency of Successes is very close to the probability of Success. This still leaves open the possibility that, during an indefinite number of trials, the actual Success frequency occasionally ventures outside any given tolerance band around that Success probability. But Borel's work killed that notion stone dead. Given any such tolerance band, there will come a time (we can't be sure when, but it will happen) after which the

actual frequency of Successes stays inside the band permanently. This is known as the Strong Law of Large Numbers.



3. Illustration of the Strong Law of Large Numbers; p is the probability of Success, the dashed lines show a tolerance band. After trial T, the actual Success frequency stays permanently within the tolerance band

This Strong Law also extends to wider circumstances: we can sum up the message of the Laws of Large Numbers by the informal phrase:

in the long run, averages rule.

In 1924, Alexander Khinchin published the wonderfully named Law of the Iterated Logarithm. Like the earlier work of Bernoulli and Laplace, when this was applied to a random quantity that arose as a sum, it could give even more precise information on how close that sum would be to its average value.

For some three hundred years, advances in the workings of probability came from a range of ad hoc methods. Then in 1933, the outstanding Russian scientist Andrey Kolmogorov used the recently developed ideas of measure theory to set the subject in a satisfactory logical framework. All the known theorems could be recast in Kolmogorov’s setting, giving a precision that was a catalyst for the developments that followed.

Kolmogorov, along with Khinchin and their student Boris Gnedenko, also greatly extended Laplace’s work on sums of random quantities. They were motivated by ways of increasing the reliability of machines used in textiles and other manufacturing industry, quality control on production lines, and the problems caused by congestion.

Kolmogorov was a superb researcher and teacher. When he died in 1987, the then Soviet president Mikhail Gorbachev rearranged his duties so as to be able to attend the funeral.

More modern times

War has frequently provoked scientific advances. The 1939 - 45 conflict boosted the development of operations research, with much of its success resting on sensible use of the ideas of probability. To maximize the probability that a supply ship would avoid being sunk by enemy submarines, a combination of data and calculation led to the conclusion that convoys were better than single ships, and large convoys better than smaller ones. When this conclusion was acted on,

losses fell dramatically. The outline of the codebreaking work in Bletchley Park is now well known: however, the importance of the use of Bayes' Rule to identify the most promising path to discover the settings of the reels on the Enigma machines is often overlooked.

In 1950, William Feller published his magnificent introductory book on probability, with further editions in 1957 and 1968. This book is my nomination for the best non-fiction book ever written. Directly and indirectly, with its mixture of intuition and rigorous argument, it led to a spectacular growth of interest in the subject. A little later, Joe Doob used the term martingale (which originally meant the gambling 'system' of doubling the stake after each loss) for a collection of random quantities where, loosely speaking, their average value at some future time is the same as the current value. He developed the main properties of martingales and closely related ideas: this work was widely useful, as it turned out that many collections of random quantities of practical interest fell within the scope of this theoretical investigation. Later we will illustrate how probability concepts have been usefully applied across a range of fields.

Many academic journals specializing in probability have been launched, some have spawned offspring, none report that they are short of material well worth publishing. The capabilities of modern computers have transformed the environment for calculating probabilities: their speed and

storage capacity have greatly increased the range of soluble problems. Earlier, most work was when probabilities were influenced by just one factor, say time or distance, and exact calculation by humans was often possible; now, complex problems where probabilities change with time, three dimensions of space, and other influences, have been successfully attacked.

Even so, it may well be that the largest influence of computers on the development of probability is through ease of communication. The language T E X has become the standard framework in which mathematics and much of science is written up. Research workers post their thoughts and ideas on the internet, scholarly articles can easily be accessed from home or office on the World Wide Web.

Chapter 4 Chance experiments

Think of any experiment with chance outcomes – buying a lottery ticket, betting on a horse race, going on a blind date, undergoing some medical treatment. We use the word distribution to specify all the possible outcomes, along with their associated probabilities. (We slipped in that word when writing about Poisson's analysis of how many rare events will happen, given a large number of opportunities.)

The 'distribution' is central to analysing the range of consequences from a chance experiment. Plainly, we need to be clear about the full extent of the possible outcomes. To give sensible values for their probabilities, we must spell out our assumptions, and hope that they are appropriate for the experiment we seek to investigate.

Discrete distributions

First, we look at circumstances where the possible outcomes can be written as a list, each outcome having its own probability. The phrase discrete distribution applies here.

The most straightforward case is when we can count the number of outcomes, and agree that they should all be taken as equally likely. The term uniform distribution is used, as the total probability is spread uniformly over the outcomes. Many experiments are expected to fit this bill – roulette, dice,

hands of cards, selecting the winning numbers in a lottery, etc. Accurate counting generates the appropriate answer.

Recall the term 'Bernoulli trials' to mean a sequence of independent experiments with a constant probability of Success each time. With a fixed number of Bernoulli trials, there is a simple formula, called the binomial distribution , that gives the respective probabilities of exactly 0, 1, 2, . . . Successes. This formula depends only on the number of trials, and the Success probability. As you run through the outcomes in order, their probabilities initially increase up to a maximum value, then fall away towards zero. (Poisson distributions also follow this pattern.)

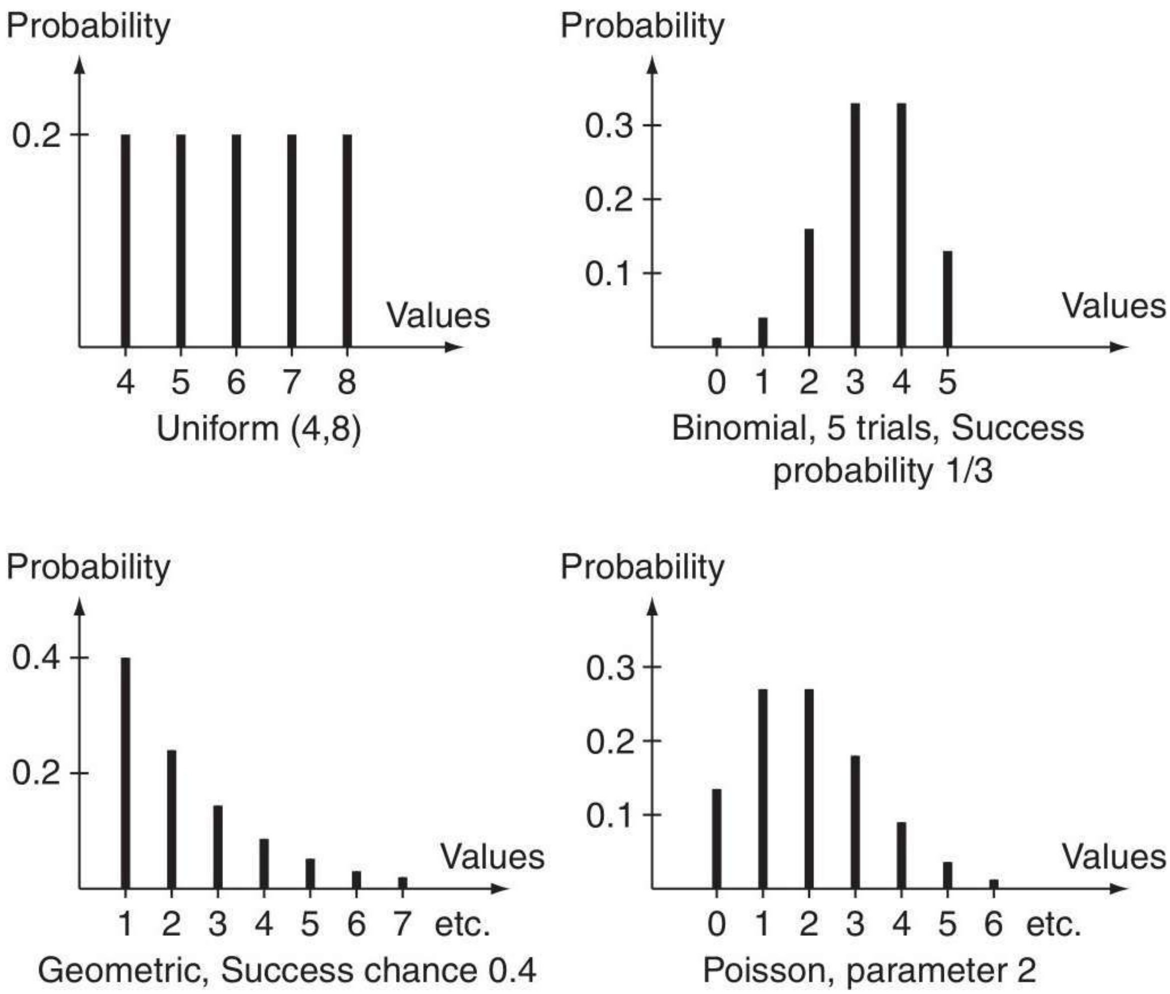
We expect a binomial distribution for the number of Sixes among twenty throws of a die; or the number of correct answers when a student guesses randomly among five choices at each of thirty questions on a multiple choice exam. But we do not expect it when asking how many Clubs a bridge player has among his thirteen cards: although each separate card has probability one quarter of being a Club, successive cards are not independent, as the chance of a Club on the next card is affected by all previous outcomes.

Always read the small print. A binomial distribution requires three conditions: a fixed number of trials, each independent of the rest, and with a constant chance of Success.

In a sequence of Bernoulli trials, what is the chance it takes exactly five goes to achieve the first Success? The only way this happens is to begin with four Fails, then have a Success; and since all trials are independent, the answer comes from multiplying the respective probabilities of these outcomes together, giving a pleasingly simple expression, the so-called geometric distribution .

The probabilities of taking exactly 1, 2, 3, . . . trials for the first Success decrease steadily. Each time, the next probability comes from multiplying the present value by the chance of one more Fail, some fixed value less than unity. Thus, whatever the chance of Success, the single most likely number of trials to achieve the first Success is always unity!

Make a leap of faith, and suppose that, in cricket, successive balls form Bernoulli trials. A bowler, who interprets 'Success' as meaning that he takes a wicket, can think optimistically: when he comes in to bowl, the single most likely time he will take his next wicket is with the next delivery. Conversely, a batsman who takes a similar view must fatalistically accept that the most likely duration of his innings is that he faces just one ball. (Even for the best batsmen, records confirm that their single most likely total score is usually zero!)



4. Some common discrete distributions

Figure 4 illustrates some of the common discrete distributions. For each possible value, the height of the vertical bar gives its probability, and the sum of all the heights is always, of course, unity.

Continuous distributions

How might we extend the classical ideas of probability to deal with the experiment of choosing a random point on a

stick of length 80cm? Here there is a *continuum* of possible outcomes, not just a list.

‘At random’ means that all individual points have the same probability. But if that common value were to exceed zero, then, by taking sufficiently many points, their total probability would exceed unity, which is impossible. Each separate point must have probability zero, and we can no longer use pictures like Figure 4. Rather than associate probabilities with individual points, we need to associate probabilities with segments, or intervals.

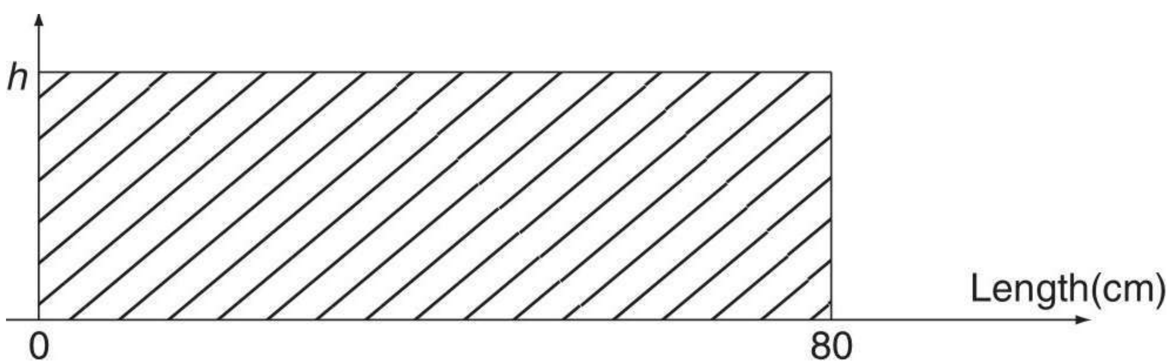
To give equal treatment along the 80cm stick, all segments having the same length must have the same probability. Imagine chopping the stick into eight equal pieces: a ‘random’ point must, by definition, fall in each with the same probability, so, for example, the segment from 20cm to 30cm must have probability 1/8.

Figure 5a shows how to proceed, using the mantra ‘Area represents probability’. The height of the horizontal line labelled h is chosen so that the shaded area beneath that line is unity, representing the fact that it is 100% certain that the random point falls *somewhere* along the interval from 0 to 80. Then Figure 5b shows how to find the probability of falling in the segment from 32cm to 52cm, by calculating the corresponding shaded area. Plainly, this is 1/4.

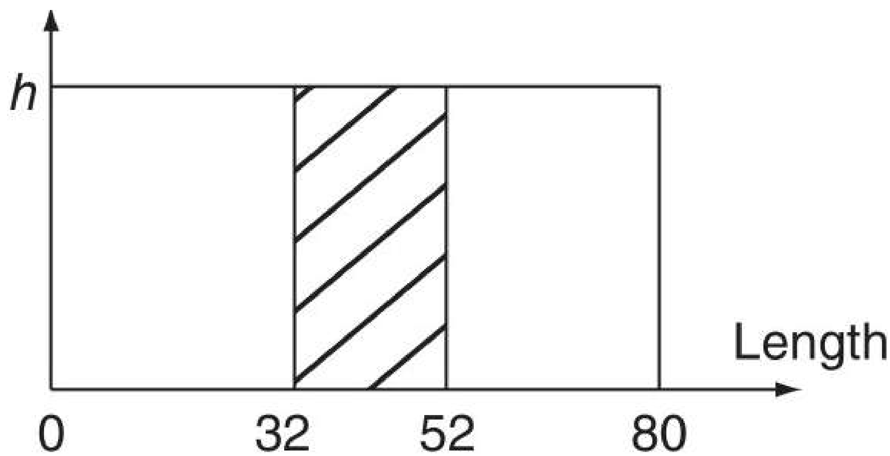
To find the probability that a randomly selected point is within 10cm of either end of the stick, or within 10cm of the

centre, we could use Figure 5c , and appeal to the Addition Law. The required probability is the sum of the three shaded areas, namely one half.

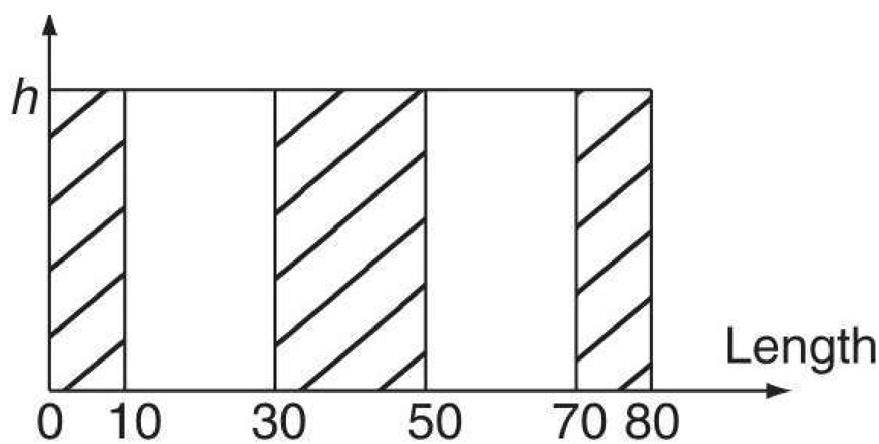
Figure 6 illustrates a similar path for other situations where the outcome takes continuous values, such as the time until the next accident on a particular stretch of motorway. We will argue below that the general shape of the curve shown is reasonable in this situation, but the main point is that the scale is chosen so that the total area above the line marked ‘Time’, but below the curve beginning at the point E, is unity, as it is 100% certain that the time to wait takes some non-negative value.



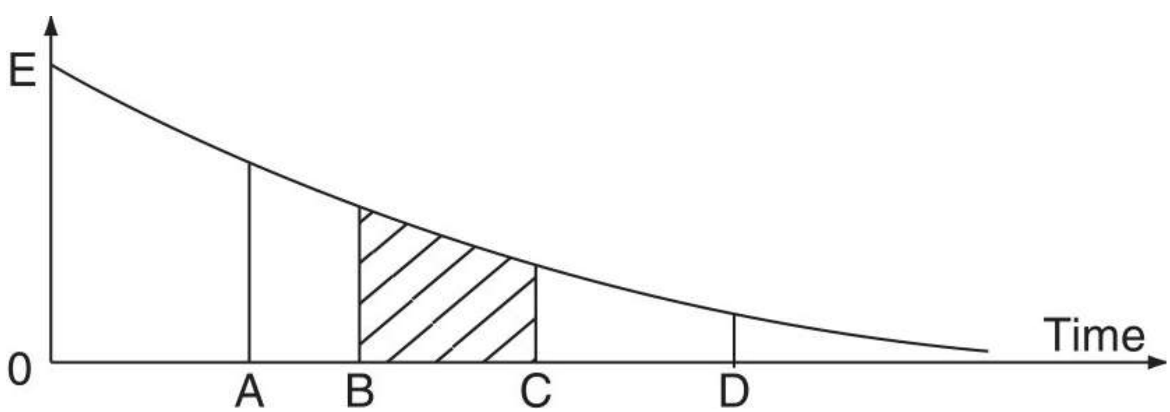
5a. The shaded area is unity



5b. The probability of falling between 32 and 52 is 1/4



5c. See text



6. A continuous distribution

The probability that the time is at least B, but no more than C, is the size of the area shaded. In a similar fashion, we can find the probability that the time to wait falls in any given interval, and then, using the Addition Law as above, the chance it falls in more complex regions.

A curve that generates probabilities in this manner is called a probability density . Now area is calculated as 'length times breadth', and the breadth of any line is zero. Hence the 'area' of either of the vertical lines at A or D in Figure 6 is zero, so both those individual points have probability zero, as before. But the density curve is higher at A than at D, so values near A are more likely than values near D. At a glance, the Figure indicates the regions of relatively low or high probability. The term continuous distribution is used.

In all such experiments, since individual points have probability zero, we can be a little slipshod: whether an interval includes both endpoints, just one of them, or neither, the probability the outcome falls in it is the same.

To qualify as a probability density, a curve must have two properties: it cannot take negative values, and the total area underneath it must be unity. This ensures that all calculations of probabilities lead to sensible values.

Many probability density functions arise often enough for them to be given names. For the experiment of selecting a random point within a given interval, the density function will be completely fiat over that interval, as in Figure 5: plainly, all segments of the same length do indeed have the same probability. Again, the term uniform distribution is used.

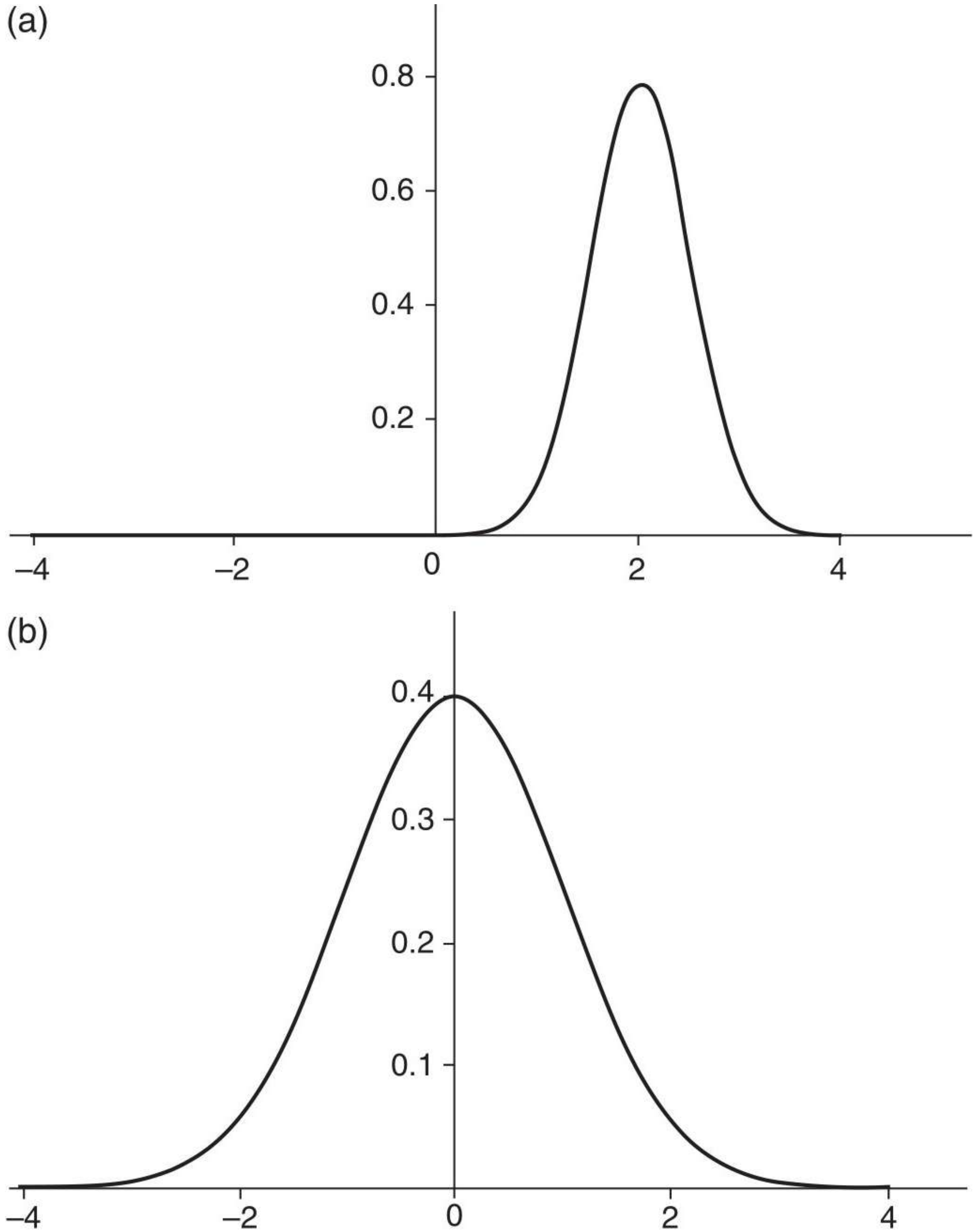
Suppose we are interested in the time to wait for some special event. For example, 210 Pb is an unstable isotope of Lead, and the claim ‘Its half-life is 22 years’ appears in physics textbooks. The meaning is that, whenever we take a lump of this substance, only half of it is unchanged after 22 years, the rest having decayed into other substances through radioactive emission.

This lump consists of a gigantic number of atoms, all acting independently. Focus on one atom: at some random time, it decays by emitting a particle. We do not know when this will be, but since half the atoms in the lump decay in 22 years, the chance that this particular atom decays within that time period is 50%. Suppose it has not decayed after five years: at that time, it is just one atom in the residual lump of 210 Pb, so the chance it decays within a further 22 years is again 50%. And if it has not decayed in the next three years, the same applies, and so on.

It turns out that the only way this can happen is when the random time until a given atom decays has what is known as an exponential distribution , whose density has the general

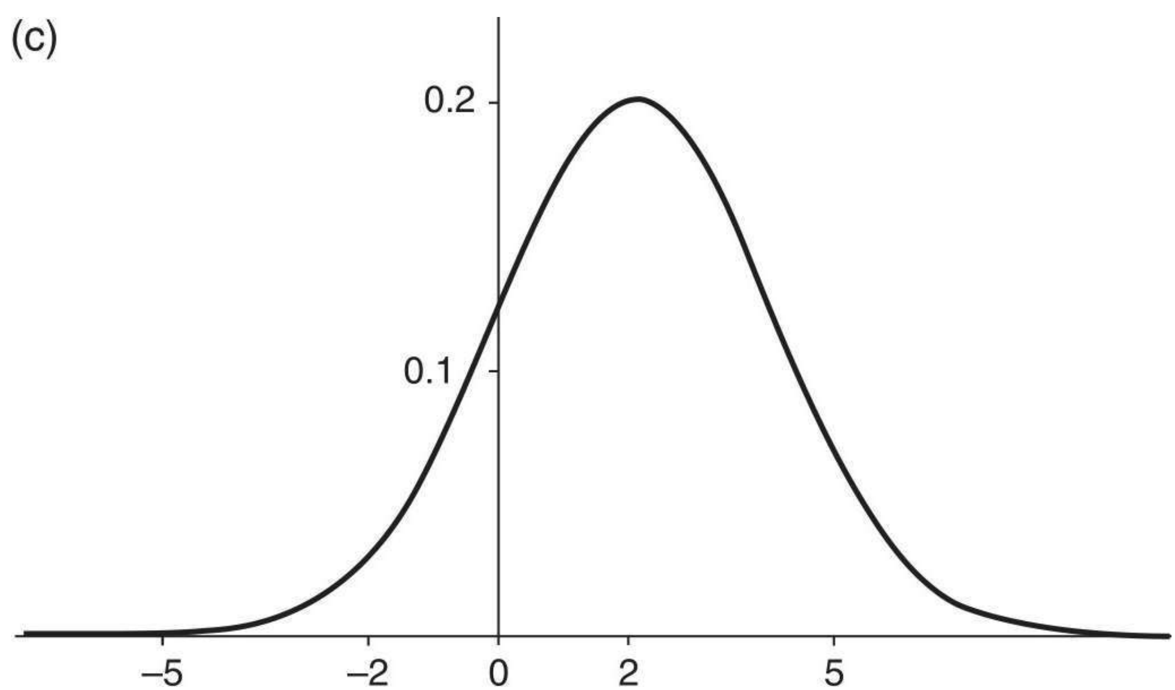
shape shown in Figure 6 , the height of the curve falling at a constant rate. A similar background applies to road accidents: if none has occurred in the past week, that seems unlikely to affect the chances of an accident in the future, so we expect the time to wait for a road accident also to follow an exponential distribution.

This distribution is intimately tied up with the Poisson distribution. Whenever things are occurring essentially at random - fiashes of lightning in a storm, spontaneous mutations in reproduction, the arrivals of some customer at the Post Office - the number of such events in a fixed time period tends to follow some Poisson distribution, while the time to wait between events has this exponential format.



7. Gaussian distributions

The most important continuous distribution is the one we have already named the Gaussian distribution . As Figure 7 illustrates, members of this family are symmetrical around a single peak and fall away rapidly towards zero, while never actually attaining that value. Two numbers tell us to which member of this family any example belongs: one number picks out the peak, the other describes the spread – small values of the spread lead to tall and narrow graphs like Figure 7a , larger values give short, fat graphs like 7c. Any probability for a member of this family can be found by using these two numbers to relate it to Figure 7b , which has its peak at zero and its measure of spread standardized at unity. Suitable tables have been widely available since de Moivre first produced them.



7. Continued

Resolving an issue

You may have noticed a problem. Provided the set of supposed outcomes is finite, or an unending list like {1, 2, 3, . . . }, then even if some members of this list turn out to have probability zero, any event whose probability is zero will never happen. However, with a continuous distribution, although each separate point has probability zero, one of them will occur when the experiment is performed! We can no longer take 'will not happen' and 'probability zero' as meaning exactly the same.

To reconcile matters, think of choosing one marble at random from a box holding a million identical marbles. We would be very surprised if we correctly guessed the outcome in advance, as the chance of doing so is only one in a million. But whichever marble gets chosen, we do not then express surprise, even though an outcome, whose probability was as small as one in a million, has occurred.

Make the box bigger - a billion marbles, a trillion - and the corresponding chance of the actual outcome can be made as close to zero as we like - but it did happen. Choosing one point at random on a continuous line is not very different from this: for any point, its chance is zero, but one of them will occur.

We will show below that, in a repeatable experiment where the chance of guessing the outcome is one in six, we can expect to perform that experiment six times in order to be

right just once. Similarly, if the chance is one in a million, we expect to take a million repetitions to guess right once. Reduce the chance of occurrence by a factor of another million, and the time we expect to wait for a correct guess gets multiplied by a million. Outcomes with really tiny probabilities do occur, but more and more rarely.

If the probability falls all the way down to zero, we can expect to wait longer than any finite time – it just won't happen! It is rational to act as though any event of probability zero, that is named in advance , will not occur.

Mean values

Knowing the distribution of the outcomes from a chance experiment, we can calculate any probability we like. But sometimes, all this detail gets in the way – we can't see the wood for the trees: so we want to pick out the main features of the distribution.

To illustrate, suppose the only outcomes possible are 2, 3, and 7, with respective probabilities 60%, 10%, and 30%. We expect that, over a hundred repetitions of this experiment, the value 2 will occur some sixty times, 3 about ten times, and 7 the remaining thirty times. The total of all these values is 120 + 30 + 210 = 360, so the average over all the one hundred outcomes is 360/100 = 3.6. This answer is just the weighted sum of the values 2, 3, and 7, the weights being their probabilities.

Whatever distribution we have, similar calculations lead to the average outcome over a large number of repetitions. 'Average' is a loose word, we prefer the term mean for the result of this calculation. There may be short cuts: if the values are uniformly distributed over some range, the mean is just midway between the two extremes; the mean number of Successes in a sequence of Bernoulli trials comes from multiplying the number of trials by the chance of Success.

When rolling a fair die, the chance of getting a Four is 1/6. So among 600 throws, we should see around 100 Fours: simple arithmetic then says that the mean gap between successive appearances of a Four is 6. It is plainly no coincidence that a chance of size 1/6 leads to a mean gap of 6. The length of any gap is just the time to wait for the next Success, so we have the pleasing result that, during a sequence of Bernoulli trials,

the mean time to wait for a Success is the reciprocal of the probability of Success.

With continuous distributions, the idea is the same, but the weighted sum is found by using the mathematical technique known as integration . For Gaussian distributions, the mean is where the peak occurs. Exponential distributions arise as the time to wait for a random event, which occurs at some characteristic overall frequency: it should be no surprise that the mean time to wait is just the reciprocal of that frequency.

The terms ‘expectation’ and ‘expected value’ are also used instead of ‘mean’ and ‘mean value’. Tossing a fair coin a dozen times, the ‘expected’ number of Heads is six, and the ‘expected’ score when throwing an ordinary fair die is 3.5. Of course, just because the expected number of Tails on a single toss is 0.5, you don’t actually expect to get half a Tail! The English language has many quirks.

Means are very friendly animals: the mean of a sum is always the sum of the means, whether or not the different components arise independently. The Law of Large Numbers says that, in the long run, means dominate: if you spend £1 on a Lottery ticket, where half that sum goes into the prize fund, then, however the prize distribution is structured, your mean return is 50p and, in the (very) long run, that is what you will get. Variability

It is also useful to have a succinct way of describing the variability of a distribution. We could calculate the difference between each value and the mean, and then find the (properly weighted) average value of these differences. But, as any trial calculation will show, this path is fruitless: the negative differences inevitably exactly cancel the positive ones, always giving a final answer of zero. But whether a difference is positive or negative, we will get a positive number when we square it. So we could use the weighted average of these squared values to assess the variability. This quantity is called the variance. If the distribution is concentrated near the mean, the variance will

be small; it will be larger when there is a reasonable chance of getting values well away from the mean.

When considering income distributions, with data in dollars, the squared data are in ‘square dollars’, whatever that might mean. Taking the square root of the variance returns us to the original measurement units, giving what is called the standard deviation .

The mean and standard deviation together often give a swift and helpful way of picking out main features of a probability distribution. And in the Gaussian case, these two numbers suffice to find any probability at all! As useful touchstones, the outcome when the distribution is Gaussian will be within one standard deviation of the mean about 68% of the time, within two standard deviations over 95% of the time, and only one time in 400 will it be more than three standard deviations away.

These figures are the basis for the guidelines offered in Chapter 1 about how close an agreement we can reasonably expect between Success probability, and the actual frequency of Success: the key is the Central Limit Theorem, which says that quantities that arise as the sum of a large number of random components are expected to follow a distribution close to the Gaussian.

In Figure 7 , showing three Gaussian density functions, the means of the graphs are at 2, 0, and 2, while their respective standard deviations are 1/2, 1, and 2.

But be warned: although the mean of a sum is always the sum of the means, the same is not generally true of either the variance or the standard deviation. If the components of the sum happen to be independent - say a casino’s profits over seven separate days in Las Vegas - then the variance of the sum will indeed be the sum of the individual variances, but otherwise it could be higher or lower.

Adding standard deviations together seldom leads anywhere sensible.

Extreme-value distributions

In several applications of probability, interest centres on the largest or the smallest of a large number of random quantities. For example, the strength of a thread or a cable rests on the properties of the weakest fibre; flood defences take account of the maximum surge that might be expected over the next hundred years; the subject of survival analysis examines what fraction of a population remains after a given time. Extreme events may occur rarely, but when they happen, the consequences can be important.

The simplest plausible model assumes there are independent random quantities, each following a particular distribution; for example, the claims made on an insurance company in each separate year. The company is interested in how big is the largest total claim it can expect to receive over the next fifty years. There is a useful mathematical result that goes a long way to answering this question:

however the claims vary over a single year, there are only three possible types of answer for the maximum claim over a large number of years. They are known as the extreme-value distributions, with the specific names of the Fréchet, the Gumbel, and the Weibull distributions. There is a sound mathematical principle that if there is some theorem about maxima, there is a corresponding result about minima. So if the item of interest is some minimum value, the same conclusion pertains.

To be able to limit the possibilities to these three families of distributions is very helpful. By estimating the mean and variance of an extreme value, and selecting whichever of them seems to fit the data best, sensible estimates of other probabilities - the chances of really extreme and devastating events - can be found.

Chapter 5 Making sense of probabilities

I will suggest how probability ideas can help in making decisions in the face of uncertainty, and also describe circumstances where misunderstandings can arise.

Odds?

Recall that probabilities can be expressed in terms of odds, and vice versa: a probability of 1/5 is the same as odds of 4 to 1 against. Unfortunately, the term 'odds' has also been usurped by the gambling community to mean something quite different - the amount the bookies will pay if your selected horse wins. So when you read that Sea The Stars won the 2009 Derby at odds of 11 to 4, that simply means that for each £4 staked on the horse, the profit, because it won, is £11. The figures '11 to 4' have no automatic relationship with the probability of winning. They depend on the bookies' subjective assessments of the horse's chances, and on how much money gamblers have staked. The term 'payout price' is a more accurate use of language for these figures like 11 to 4, but, regrettably, we have to accept the common usage of 'odds' in this gambling context.

A payout price is termed fair if it gives no monetary advantage to either party, i.e. the mean value of the gamble is zero. The fair payout price for correctly picking the suit of a card selected from a well-shuffled deck is 3 to 1, as those are the exact odds against a correct guess.

Commercial gambles are not fair, in this sense, as they could never operate without a house advantage. For roulette in a UK casino, when all 37 outcomes are equally likely, the payout price for betting on a single number is only 35 to 1, not the 36 to 1 that would be fair. So the mean return on a bet of £ 37 is £ 36, giving a house advantage - the percentage of any bet it expects to win - of 1/37, about 2.7%.

This advantage is the same for most of the available bets in roulette: whether you are betting on pairs of numbers, triples, groups of four, or six, or twelve, for every £ 37 you stake, your mean return is always £ 36. But in Las Vegas, the standard house advantage is bigger, because of an additional slot, double zero, giving 38 outcomes - with the same payout prices as in the UK. The mean return on $38 is generally $36, a house advantage of 2/38, or 5.3%.

A different way to bet on horse races is through a Tote or pari-mutuel system. Here, the money bet on all the horses is pooled, and a fixed proportion - 80% or so is common - is shared among those who backed the winner, in a size proportional to their stakes. The Tote advantage is then 20%, whichever horse wins.

The size of the bookies' advantage in a horse race depends on which horse happens to win. Although bookies may make a loss or profit on any single race, recent data tell a sobering story: at a payout price of 6 to 4 on, punters

should expect to lose about 10% of their stake; at a price of 5 to 1, expect to lose about 13%; at 10 to 1 the mean loss figure is over 23%, and if you speculate on horses priced at 50 to 1, expect to lose about two-thirds of your money.

This phenomenon is known as the favourite-longshot bias . Punters lose their money more slowly from bets on the more favoured horses than if they are attracted by large payout prices. Bookmakers were delighted when Mon Mome won the Grand National in 2009 at a price of 100 to 1.

Absolute risk, or relative risk?

Suppose that, among a particular group of people, the chance of developing colorectal cancer over the next five years is quoted as one in a thousand. We expect about ten among 10,000 people to develop the cancer. A new drug would reduce the chance to one in two thousand: then only about five among 10,000 would succumb if the new drug is used. The drug company could headline its press release 'Risk of cancer cut by 50%' . And that is accurate: for each person, the risk would be halved.

This approach describes a reduction in relative risk , and is often criticized as putting too favourable an interpretation on the data. For, suppose the initial risk had been one in ten million: cutting it by 50% leads to a new risk of one in twenty million, but in either instance, the risk is so small that among 10,000 people, we would expect pretty much the same number of cases - zero. Despite the

risk being halved, the drug would hardly ever make a difference.

But suppose members of this group had a much higher chance, say 40%, of developing the cancer. A drug reducing the chance to 20% would qualify for the same headline, and would correctly be hailed as a major breakthrough, as among 10,000 people, fully 2,000 fewer would develop the cancer.

Rather than focus on the relative risk, it is usually more meaningful to look at the change in absolute risk. In the first case above, the absolute risk changes from 0.1% to 0.05%, so the drug is 0.05%; in the second case, the drop is a minuscule 0.000005%, while with the final figures, the drop is an impressive 20%.

A sensible way to proceed is to state the mean number of patients who should take the drug in order to prevent one case of the disease - the Number Needed to Treat , or NNT. The fewer the better, and this number is just the reciprocal of the change in absolute risk. In the examples above, the respective NNTs are two thousand, twenty million, and five.

Treating twenty million people to prevent one case of a disease is hard to justify. The NNT, along with knowledge of the treatment costs and the severity of the impact of the disease, allows us to make sensible decisions about allocating health care resources.

Combining tiny probabilities

How likely is it that at least one among an enormous number of events, each having a tiny probability, will occur? This can be the pertinent question when considering the probability of a catastrophe. Complex systems or machinery may fail if any one of a myriad of components fails; will two aircraft collide, or might a nuclear power station suffer a meltdown? The so-called Borel-Cantelli Lemmas give some pointers. These mathematical results show that, in many circumstances, the key quantity is the sum of all those tiny probabilities: if it grows without bound, catastrophe is certain.

One consequence is that we can never be satisfied with current safety standards. It is essential to continue to make improvements.

For, no matter how high our standards, there is some non-zero probability of failure during any given month: and however small this value, if it remains unchanged (or even if it decreases too slowly) the sum over many months will grow indefinitely large, and disaster will occur sometime.

A sound programme of continual improvement does not guarantee that disaster will be averted: but to be ever satisfied with the status quo is to invite doom.

Some misunderstandings

(a) When a doctor tells a patient that there is a 30% chance that a particular medication will have unpleasant side effects, he means that he expects about 30% of patients on

this drug to suffer. However, the patient may believe that these effects will arise on about 30% of the occasions on which she takes the drug. The doctor is thinking about all the patients he sees, the patient about all the times she takes pills – their reference classes are different.

(b) How does the public interpret the claim ‘There is a 30% chance of rain in Chicago tomorrow’ from a TV weather forecaster? The forecasters expect their audience to make a frequency interpretation, i.e. that, in the long run, rain would fall next day on 30% of the occasions when the weather conditions were similar to those now seen.

But when questioned, even among those viewers who were happy with the phrase ‘a 30% chance’, there was a spread of beliefs. Some felt that they were being told that rain would fall over 30% of the city’s area; others that it would rain in Chicago for 30% of the day; and some believed that 30% of meteorologists expected it to rain! A few thought that it would definitely rain, with the 30% figure indicating the rain’s intensity. There were many mismatches between the event the forecasters were referring to, and the event viewers were thinking about.

(c) If as few as twenty-three people chosen at random gather together, it is more likely than not that two of them share a birthday. When people meet this fact for the first time, they are normally surprised, but usually become convinced when proper counting shows this claim to be true.

However, a minority remain unconvinced, because they mistakenly think they have been told that, if they and twenty-two others gather together, it is more likely than not that one of the others shares their birthday. Listen carefully!

(d) Suppose that a coin, accepted as fair, shows Tails on nine consecutive tosses. Some will claim that the next toss is almost certain to be Heads, perhaps by invoking some 'Law of Averages' that requires Heads to eat into this excess of Tails immediately. No such Law exists. The Law of Large Numbers does imply equal proportions of Heads and Tails, but only in the long run: any sequence of nine Tails is diluted by the thousands of tosses before and after.

Alternatively, some will (correctly) note that the chance of ten consecutive Tails is less than one in a thousand; so if they see nine consecutive Tails, they may then 'deduce' that Tails next time is highly unlikely. But that is false logic: if ever we have nine Tails in a row, a tenth Tail will happen half the time. This confusion of the absolute probability of an event, with its conditional probability, given a certain background, famously arose in 1996: jockey Frankie Dettori rode the winners of the first six races at Ascot, and since no-one had ever won all seven races in a day, it 'must' be virtually impossible for Dettori to win the last race. But he did. Very few people will ride the first six winners in a seven race meeting, but when someone does so, he might well also win the last.

Ask yourself: am I assessing the absolute chance of twenty things happening, or just the conditional chance of the twentieth, given that the first nineteen have happened?

(e) Newspapers are often produced under great time pressure, so it is not surprising that some articles contain nonsense. But here are three reports that should have been spiked.

When all six eggs in a box bought in a supermarket were double-yolked, it was claimed that a truly astonishing event had happened. Only one egg in a thousand has this property, so the chance of getting a boxful is this tiny fraction, multiplied together six times. The resulting number is so small that if you opened one box every second, you would expect to take over thirty billion years to come across a box containing only double-yolked eggs!

But that calculation makes sense only if all the eggs in a box are chosen independently from a vast collection of eggs, one in a thousand of which is double-yolked. This doesn’t happen. Eggs are sorted by size before being boxed. Some boxes are even labelled as containing only double-yolked eggs . . . .

Allegations about the private life of an England soccer captain surfaced. A reporter offered figures for the ‘likelihood’ of each of four possible actions by the team manager:

(a) expel him from the playing squad - 1/10

(b) retain in the squad, but invite him to resign - 3/10

(c) retain in the squad, but remove the captaincy - 6/10

(d) take no action - 8/10

Any single one of these four estimated probabilities is plausible. But since they relate to mutually exclusive outcomes, their sum must not exceed unity: however, these 'probabilities' add up to 1.8.

Thirdly, it was reported that, among people who had won at least £50,000 on the National Lottery, the names turning up most often included John, David, Michael, Margaret, Susan, and Patricia. So far so good: but it was absurd to claim that therefore you should try to include people with those names in your Lottery syndicate!

Describing ignorance

Drawing one card from a well-shuffled pack, I expect Black and Red to be equally likely, and I confidently attach the figure of '1/2' to the chance of Red. In a slightly convoluted way, I could say that the distribution I attach to the chance of getting Red allocates 100% to the quantity '1/2' - my confidence is shown by the choice of 100%. If I am sure that a probability is some definite figure, I attach probability 100% to that figure.

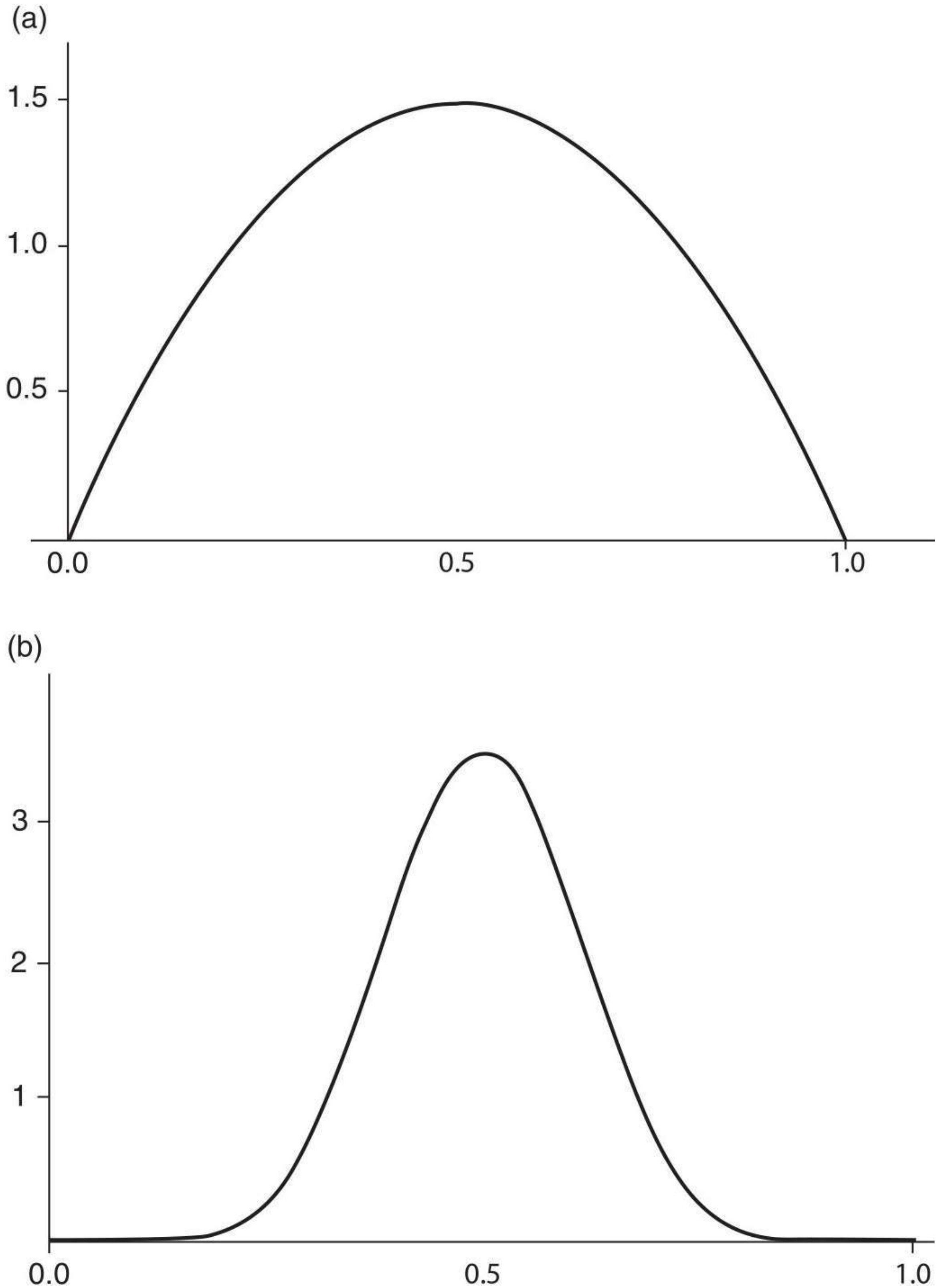
But frequently, I cannot select a single figure: my best estimate of the chance my train will miss its connection may be 3/4, but 2/3 and 4/5 could be nearly as plausible, and even values close to the extremes of zero and unity are not completely ruled out. I can use a continuous distribution over the range from zero to unity to describe my feelings about this unknown probability.

Total ignorance of a probability - very rare circumstances - would be described by using the continuous uniform distribution of Figure 5. More often, there is some intermediate figure which is your best single guess at the probability, and your instinct about both higher and lower values is that their chances fall away towards zero. Figure 8 shows a collection of graphs from what is called the beta family of distributions that have this property.

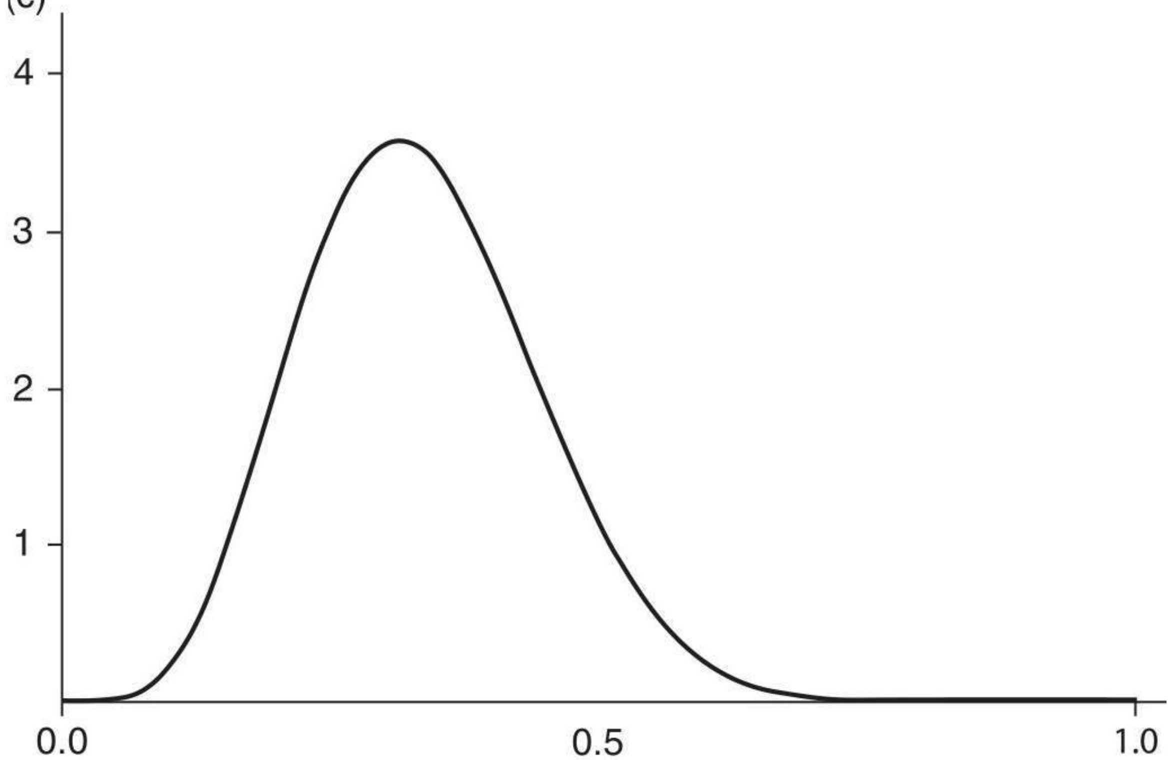
Figure 8a indicates that we are pretty ignorant of the value of the probability: we attach highest belief to values near 1/2, but values as small as 1/5 or as high as 4/5 are still quite possible; with Figure 8b , we are much more confident that the value is close to 1/2, while still not ruling out the extremes. With 8c, our highest belief is for values close to 1/3, while with 8d the values are very strongly congregated near 2/3, with very little expectation that the value is below 1/2.

How far will my car travel on 10 litres of petrol? When warmed up, and driven at a steady speed, I expect to get

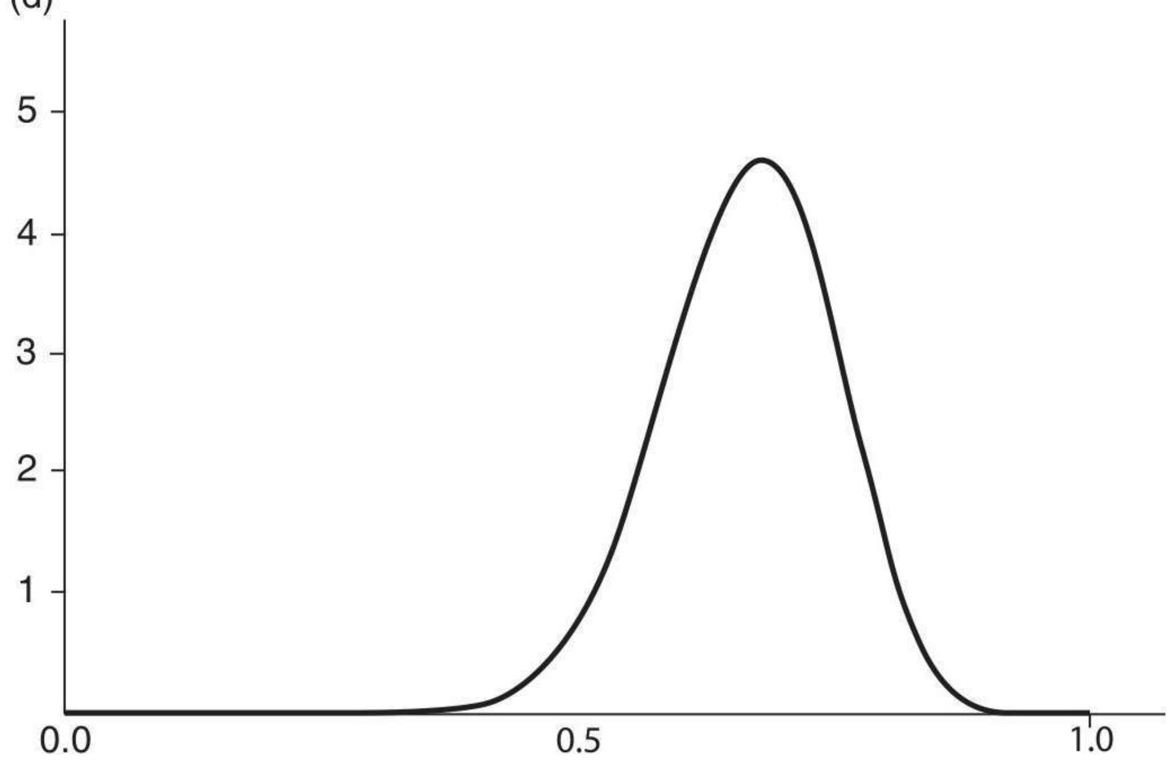
about 90 miles, but if I make short journeys over a couple of weeks, it will be more like 60 miles. In either case, the distance will have some uncertainty, expressed via some continuous distribution. To get a handle on what kind of distribution, imagine splitting the petrol into small cups of size 10cc. There will be 1,000 such cups, and the total distance covered will be the sum of the distances achieved using these 1,000 components. Recall the Central Limit Theorem, which says that the sum of a large number of random quantities will tend to follow a Gaussian distribution.


(a)
1.5
1.0
0.5
0.0
0.5
1.0
(b)
3
2
1
0.0
0.5
1.0

(c)



(d)



For steady motorway travel, I would select a Gaussian distribution centred on 90, sharply peaked to show low variance. For sporadic travel within town, I would also use a Gaussian distribution, but centred on 60 and with a wider spread to demonstrate the greater uncertainty.

Utility

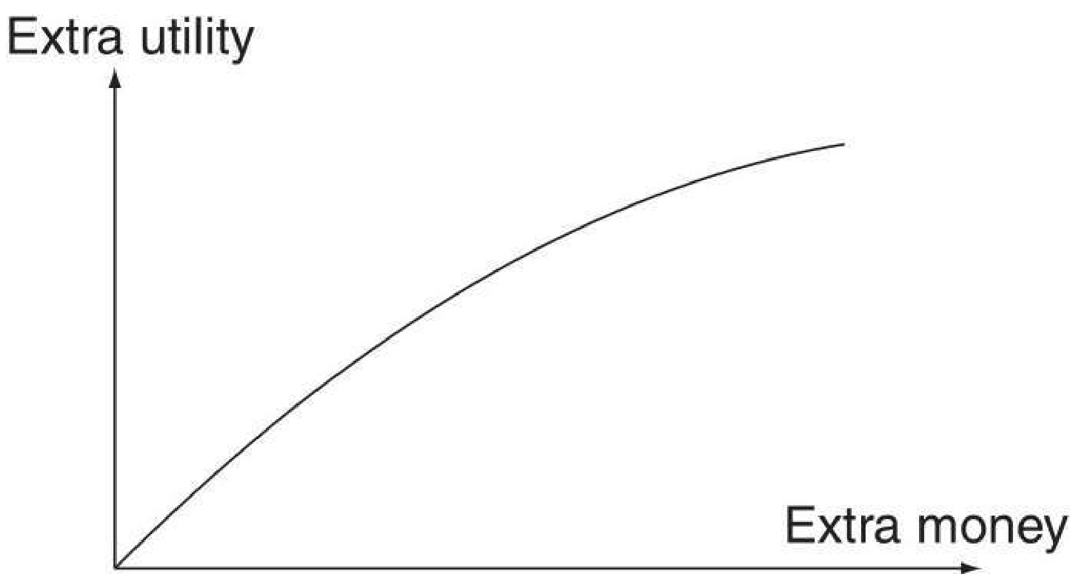
Your fairy godmother makes you a once only offer. Either she will give you £1, or she will toss a coin and if you call Heads or Tails correctly, she will give you £10, otherwise you get nothing. What choice would you make?

The alternatives are £1 for certain, or a fifty-fifty gamble with a mean payout of £5. Nearly everybody prefers the latter. But scale the money up by a factor of one million: overwhelmingly, preferences change. To have one million pounds for certain is far more attractive than being equally likely to have zero, or ten million pounds. The concept of utility lies behind this difference.

For small sums of money, having twice as much usually is worth twice as much, but if one million pounds would generate a certain level of pleasure for you and your family, double that amount would not lead to double the pleasure. In whatever units you choose to measure the 'worth' of a monetary amount, the usual shape of the relationship will follow that shown in Figure 9 : the graph always rises,

initially like a straight line, but then steadily more slowly.

'Utility' explains why a householder and an insurance company can agree that a sum of £250 is a reasonable annual premium to insure a house worth £250,000 against perils such as fire, subsidence, orflooding. The substantial reserves of the insurance company mean that, on any individual house, it can act as though the utility is identical to the relatively small sums involved. So long as the chance that it will pay out in any year is less than one in a thousand, its expected value for this transaction is positive, it will make a profit. On the other hand, the uninsured householder would face the enormous negative utility of finding £250,000 if she lost her house to one of those perils, so voluntarily surrendering just £250 to remove that possibility is a good deal to her too.



9. The general shape of a utility curve

Taking out insurance against breakdown of televisions, microwave ovens, and so on is almost always a bad idea. The sums are much smaller, utility and money are essentially the same, and the company will charge a premium large enough for them to expect a profit. Instead of buying this expensive insurance, build up a repair fund by placing the hypothetical premium into your bank account. Very few people will regret that move.

If you can successfully construct your own utility function, you can use it to help make a choice between the available actions under conditions of uncertainty. For each action, calculate the expected utility of the outcomes, i.e. weight the utilities by their respective chances. Then select that action for which this expected utility is as large as possible.

That is the probabilist’s universal recipe for making the best of the available choices, whatever the circumstances.

Chapter 6 Games people play

Many recreational games combine skill and chance. Skill you can work on, chance is, well, a matter of luck. For all the ‘games’ discussed here, it is easy to persuade yourself that there is some finite list of outcomes, all equally likely. Thus in this chapter, unless explicitly stated otherwise, we will use the classical approach to finding probabilities: count the number of possible outcomes, and the probability of any event is taken as the proportion of those outcomes where the event happens.

My aim is to show how the ideas of probability can help a player make good decisions under conditions of uncertainty. An understanding of probability can also add to the enjoyment or entertainment of spectators.

Lot teries

A common lottery format is that known as 6/49, as in the UK National Lottery. Here 49 rubber spheres, painted with different numbers, are whirled around a plastic tub, six are chosen at random. Gamblers pay £1 to select six numbers, and win a prize if their selection contains at least three of those winning numbers. But since only 50% of the takings go into the prize money, the mean return to Lottery players is far less than in casinos, or at the racetrack.

The main attraction is the prospect, however remote, of an enormous prize - one UK ticket has won over £22 million, and prizes in the USA have exceeded $300 million. Counting tells us that the probability of winning a share of the top prize from one ticket, by matching all the winning numbers, is about one in 14 million in the UK, less than one in 116 million in the Euromillions game, and around one in 176 million in the USA Mega Millions game.

To appreciate just how tiny these chances are, fix on the UK game. Figures show that the probability of death within one year for a random 40-year-old man is around one in a thousand. So the chance of his death within a day is about one in 365,000, within an hour it is about one in 9 million, so to get down to one in 14 million we are talking about the chance he dies within the next 35 minutes! For the Mega Millions game, under the same assumptions, the chance his ticket wins a Jackpot share is comparable to his chance of death within the next three minutes.

Despite the low return and forbidding odds, 'utility' gives a rational explanation for buying tickets. In exchange for £1, you will get back 50p on average anyway, and the other 50p buys you the right to dream about your future luxurious lifestyle, your philanthropy, and the possible envy of people like me who assured you it was a waste of money. These rights surely have some utility.

We shall assume that future lottery draws are independent of past results – an inanimate rubber sphere cannot ‘remember’ whether it is ‘due’ to be chosen. Short of cheating, there is no way of changing the chance of winning a prize. But you can influence the size of any prize, in those lotteries where a fixed proportion of the prize fund is shared out among the winners at each tier. There is an opportunity to exercise a little skill.

It arises because certain numbers, typically low (birth dates) and odd, are chosen more often than others, and because many lottery players spread their choices evenly over the ticket, perhaps in the mistaken belief that doing so means that they are selecting ‘at random’. In consequence, combinations with several high numbers, or with numbers clustered together, or on the edge, are chosen less often. If you can identify the sort of choices other players are making, and do something different, your winning chance is not affected, but if you do win, you will win more than average.

Beware of trying to be too crafty, like selecting {1, 2, 3, 4, 5, 6}, or the winning numbers in the last draw, on the grounds that ‘Nobody else will think of that’. They will. When the UK Lottery began, about 10,000 people were choosing the first six numbers. In September 2009, the winning numbers were identical in two successive draws in the Bulgarian Lottery: no-one chose them the first time, but 18 did so the second time.

Provided that other players continue to mark their tickets much as they have in the past, the following procedure, for UK-type 6/49 lotteries, will help you. Take an ordinary deck of 52 playing cards and discard three of them. Identify the remaining cards with the numbers 1 to 49, shuffle well, and select six cards. This is a way of choosing six numbers completely at random. Human beings cannot make such a selection unaided, they need this sort of auxiliary help.

And use these six numbers, provided that

(a) they total at least 177 (to give a bias to high numbers), and

(b) when marked on the ticket, they fall into two, three, four, or five clusters, and

(c) three, four, or five of them are on the outside border of the ticket, and

(d) they do not form any obvious pattern on the ticket.

If any of these conditions fail, return the six cards to the pack, shuffie it thoroughly, and repeat this sequence.

If you follow this recipe, you should still expect to lose money - the overall payback of only 50% of money received is hard to overcome. But you are less likely to have to share any Jackpot with the world and his wife.

TV games

Golden Balls was first aired in 2007. The last two players are faced with eleven spheres (the Golden Balls), some of which are worth money, the rest (the Killers) are worth zero. The players select five of these spheres to generate a potential prize fund; any Killer chosen reduces the value of any previously selected Ball to one tenth of its current value. Thus two Killers chosen after a £50,000 Ball make it worth just £500.

All the spheres are outwardly identical, so the players are choosing completely at random. There are 462 ways to choose five objects from eleven, so the chance of picking the five most valuable Balls is 1/462. In the first 288 shows, this occurred just once.

Take a Ball nominally worth £1,000: even ignoring the Killers, the chance of selecting it is only 5/11, so its real mean value is £455. Any Killers will reduce this sum even further – with three Killers, its mean value can be calculated as £255.

When the five Balls have been chosen, and the actual prize fund is known, each player makes a private decision, whether to Split the fund with the other player, or seek to Steal all of it. They reveal their choices simultaneously: if both Split, they share the fund, if just one of them Steals, that player gets the entire fund, if both Steal, neither get anything.

This scenario is well known in Game Theory, under the title 'The Prisoners' Dilemma'. Suppose your opponent chooses Split: then you are better off if you Steal. And if the opponent chooses Steal, you won't get anything anyway. So whatever choice the other makes, you can argue that choosing Steal will never lose. Frequently, both select Steal, and the only winner is the TV production company who pay out zero.

Versions of Deal or No Deal have been shown in over seventy countries. In the UK, there are 22 sealed boxes containing different amounts ranging from 1p to £250,000. The boxes are allocated randomly to 22 players, one of whom, Amy, will play that day. Her own box remains closed until the end. She first selects five other boxes, whose contents are revealed. A banker then offers a sum of money in exchange for the amount in her box. To accept this, she says 'Deal', ending the game, while the words 'No Deal' reject this offer. If the game continues, more boxes are opened, a new offer is made, and so on.

At the time of any offer, the exact amounts still in play are known, so their mean is easily calculated. In the early rounds, the banker's offer is normally far less than this amount, but Amy must have her utility function firmly in sight: if she strongly desires £5,000, and the offer is £5,400, she could rationally accept, even if the mean amount in the boxes left is over £20,000 - she might end up with 1p if she hangs on.

One time in 22, Amy will own the box with the top prize, but she will win that amount far less often. Utilities give a convincing explanation. At the final decision point, two boxes remain, one with £250,000 and the other with maybe £2. If the banker offers £80,000, even though this is well below the mean value of £125,001, only the bravest or richest Amy will reject it. A bird in the hand . . . .

Provided the banker always offers less than the mean amount in the remaining boxes, the Law of Large Numbers ensures that, in the long run, contestants who Deal take home less than the amount in their box. So a real bank, that did pay out the offer but also received the amount in the box when the contestant Deals, would make a long-run profit.

The Colour of Money was billed as the most stressful show on TV, yet it survived only a few episodes in 2009. But it does give a splendid opportunity to illustrate uses of the Addition and Multiplication Laws in finding a probability.

The sums £1,000, £2,000, . . . , £20,000 were randomly allocated to twenty boxes of different colours, and the player, Paula, sought to reach some pre-assigned target, say £64,000. To do so, she could select up to ten boxes, one at a time. If she (unknowingly) chose the £14,000 box, the amounts £1,000, £2,000, . . . up to £14,000 would appear in that order at a stately pace: she could call Stop at any stage. If she made that call in time, she banked the amount last showing, but if she waited too long, she banked nothing.

If, after ten boxes, she had not reached the target, she won zero. What tactics should she use?

Colour apart, all the boxes are identical, so Paula makes a completely random selection from those left in each round. In her last round, with eleven boxes left, her strategy will be obvious: for example, if she needs a further £9,000 to hit her target and exactly six boxes are worth £9,000 or more, she will hope to call Stop when £9,000 is shown, and her winning chance is 6/11. But what should she do in earlier rounds?

Perhaps the twelve boxes left with two rounds to go contain (in units of £1,000) the amounts 1, 4, 5, 6, 9, 10, 12, 13, 15, 17, 19, 20, and she requires another £15,000. It makes no sense to call Stop when she sees £7,000; if that figure ever appears, she knows that her box contains at least £9,000, so she could Stop at that sum, plainly better tactics. She can restrict her options to selecting from the twelve values in the boxes. The same argument also applies at the earlier rounds – her best call of Stop will always be at a value corresponding to one of the remaining boxes.

If Paula does intend to Stop at £9,000 here, she can argue: 'Eight of the twelve boxes have at least that amount, so my chance of success is 8/12. And if I do succeed, I'll need just £6,000 in the final round, and then eight of the last eleven boxes will work. The Multiplication Law tells me that chance of both of these is (8/12)*(8/11) = 64/132. Also

four boxes have less than £9,000, so the chance I bank nothing first go is 4/12; I then need £15,000 from the last box, with chance 4/11. By the Multiplication Law again, the chance this path will work is (4/12)*(4/11) = 16/132. These ways of winning are disjoint, so the Addition Law gives the overall chance of success as 80/132.'

She can make a similar analysis for her other choices, such as going for £6,000, or £12,000. I invite you to do the sums – the Appendix describes her best choice.

In the planning stages of this show, the idea of using an expert mathematician to offer running advice was mooted. Paula could suggest she will try to call Stop at £8,000, the expert might say 'Not a bad choice. If you do that, you've got a 75% chance of winning the money. But if you plan to Stop at £11,000, your chance goes up to 80%.'

You can well imagine what could happen! Everything the expert said was correct, Paula changed her call – and failed to win the money, while her original instinct would have worked. Some tabloid newspaper would surely scream 'Maths Boffin Robs Army Hero's Widow of £64,000'.

All of us who investigate the maths of TV game shows are relieved that no such mathematical advice was ever given on this show!

Card games

The Law of Large Numbers means that you will receive your fair share of good or bad cards in the long run, so differences in skill levels will tell eventually. We look at three popular games. In Blackjack, the house must follow fixed rules about when to draw cards, the player can do what he likes. Until Edward Thorp started winning significant sums, casinos believed that no system could beat their built-in advantage. Their logic had a fatal flaw: although they could expect to win 1- 2% of stakes with a full stack of six or eight decks of cards, after a few deals the odds might shift in favour of the player. The casinos had omitted to use the conditional probabilities based on what cards remained unused, rather than rely on the probabilities calculated for a full stack.

Thorp developed a way of keeping track of which cards were left in the stack. When there is a high proportion of high value cards remaining, it becomes more likely that the rules compel the house to draw a card that leads to a losing total of above twenty-one: in the same circumstances, the player can opt not to draw. Thorp would bet the minimum amount so long as the stack had a low or average proportion of high value cards, then larger amounts should the stack composition shift in his favour. Simple, but effective.

When the stack composition does give an advantage to the player, how much should he bet? John Kelly had answered precisely that question a few years before Thorp’s analysis: he should bet that proportion of his capital that is equal to

the size of his advantage. This choice maximizes the rate at which his capital will grow.

For example, suppose he has £1,000 and the game is slightly in his favour - his winning chance is 51%, his losing chance is 49%. His advantage is 2%, so he bets 2% of his current capital, i.e. £20. Next time, he will have either £980 or £1,020, so if his advantage remains at 2%, his bet will be either £19.60 or £20.40, according to which outcome pertained. If he were too greedy - betting 10% of his capital when Kelly indicates just 2% - then he would eventually be ruined, despite his advantage. His capital is finite, and the stake would be too high to stand the inevitable losing streak.

Casinos take steps to identify and ban proficient card counters. No finer tribute to the power of understanding probability has ever been paid.

We noted that Bayes' Rule is the proper way to see how pieces of evidence should change our beliefs about Guilt or Innocence in a court case. In card games like Whist or Bridge, using this Rule can improve your chances of making the best decisions during play. For convenience, I retain the legal vocabulary, and use the word Guilty to mean that a particular opponent holds certain cards, say both King and Queen of Hearts, while Innocent means that she holds at most one of those cards.

By counting, we can find the proportion of all possible deals where she holds both cards, to give an initial assessment of the probability of ‘Guilt’. It turns out best to convert this probability into its equivalent odds, in the standard manner. As this calculation is made at the outset, we say that we have found the Prior odds (of Guilt).

As cards are played, relevant Evidence emerges - perhaps she plays the King of Hearts on a trick. To see how such Evidence affects the odds of Guilt, a quantity termed the Likelihood Ratio is found: first, assess the probability of the Evidence assuming Guilt (she holds both King and Queen), then find its probability assuming Innocence (she has at most one of them). The Likelihood Ratio is just the ratio of the first to the second.

We can now deduce the Posterior odds, i.e. the odds of Guilt, taking account of this Evidence, using Bayes’ Rule which says

Posterior odds = Prior odds ≈ Likelihood Ratio.

Its format is plainly sensible: the bigger the Likelihood Ratio (i.e. the more likely is the Evidence when the opponent is Guilty), the more the odds of Guilt increase - but this Rule tells you precisely how much the Evidence affects the chances of Guilt.

To see it in operation, consider a realistic situation: our opponent either holds just the King and Queen (Guilt), or

she has the King only (Innocence); the Prior odds are that those alternatives are just about equally likely. If she is Guilty you do best to play the Ace, if she is Innocent you should play some other card. Evidence now appears - she plays the King.

Without the Evidence, you must guess, and you will make the winning play half the time. With Innocence (she has King alone), the probability of the Evidence (she played the King) is 100%; but with Guilt (she had both King and Queen), she might equally well have played the Queen rather than the King that you saw, so the probability of the Evidence is only 50%. Their ratio is one half, so the Rule tells you that the Posterior odds are one half, i.e. she is twice as likely to be Innocent as Guilty - she is twice as likely to have the King alone. Not playing the Ace is the right decision two thirds of the time.

If, by this proper use of probability, you will make the winning play two thirds of the time, rather than just half the time, you should expect to do much better. You cannot guarantee to make the winning play, but you can improve your chances of doing so.

Bridge players refer to this scenario as the Principle of Restricted Choice - if the opponent had King alone she had to play it, with both King and Queen she had a choice. The fact that she did play the King shifts the odds towards her having to do so.

Today, the most popular form of Poker is Texas Hold’Em. Each player is dealt two cards and seeks to make the best poker hand possible from her own cards, and five communal cards that are dealt face up later. Which of the following hands is best, in the sense of being more likely to beat either of the other two when those communal cards are dealt?

Hand A: Two of Clubs, Two of Spades

Hand B: Ace of Spades, King of Diamonds

Hand C: Jack and Ten of Hearts.

Trick question, of course: after careful counting, it turns out that Hand A will beat B about 52% of the time, B beats C 59% of the time, while the chance C beats A is around 53%. So you would rather hold A than B, and rather hold B than C, but also you prefer C to A! Your chance of winning exceeds 50% if you let your opponent pick any of the three hands, provided you may then choose either of the others for yourself.

There is far more to poker than facility with probabilities. You must make judgements about what cards your opponents are likely to hold, and when you might bluff. But sometimes probability is very useful. Suppose the pot has 50 chips, and one communal card remains to be dealt. You can see that, if this final card is a Spade, you will make a Flush, which must win; if it is not a Spade, an opponent will win. Should you bet more chips to remain in the game?

Ignore how much you have already put in the pot. All that matters is the future. You can see six cards - two in your hand, four communal cards on the table. Of the 46 unknown cards, nine are Spades that give you victory, the rest lead to defeat. With 50 chips already in the pot, is it worth paying 10 more to see the final card dealt? 20 more?

By working out the mean profit (or loss) if you must pay x chips to see the final card, find the cut-off value of x that will, on average, give a profit. The Appendix gives the answer.

Chapter 7 Applications in science, medicine, and operations research

We may assess or interpret probabilities in different ways according to the context. But, as David Hand wrote in his Statistics: A Very Short Introduction , ‘. . . the calculus is the same’ , i.e. how probabilities are manipulated does not change.

Keep in mind the central ideas of the subject: the Addition and Multiplication Laws; independence; the Laws of Large Numbers linking frequencies to objective probabilities; Gaussian distributions when summing random quantities; other frequently arising distributions; means and variances as useful summaries.

We may not expect our knowledge of the relevant probabilities to have the precision available for the examples in the previous chapter, but an approximate answer to the right question can be a reliable guide to good decisions. As statistician George Box said: ‘All models are wrong, but some are useful.’

The next two chapters illustrate applications, loosely grouped under the chapter titles.

Brownian motion, and random walks

In 1827, the botanist Robert Brown observed that pollen particles suspended in liquid move around, apparently at random. Nearly eighty years later, Albert Einstein gave an explanation: the particles were constantly being buffeted by the molecules in the liquid. This movement is, of course, in three dimensions, but to build a satisfactory model, we first consider movement just along a straight line.

Suppose that each step is a jump of some fixed distance, sometimes left and sometimes right, independently each time. This notion is termed a random walk . The position after many jumps depends only on the difference between the numbers of jumps in each direction; the mean and variance of the distance from the start point are proportional to the number of jumps made.

Make a delicate computation: over a fixed time period, increase the frequency of the jumps, and decrease the distance jumped. With the correct balance between these two factors, the limit becomes continuous motion, the random distance moved having (by the Central Limit Theorem) a Gaussian distribution whose mean and variance are both proportional to the length of the time period. If movements left or right are equally likely, the mean will be zero.

Einstein’ s explanation for Brown’ s observations is that particles move in three dimensions, movement in each dimension following a Gaussian law for the reasons given. He made predictions about how atoms and molecules behave,

provoking experiments that removed any lingering doubts about their existence.

The term 'Brownian motion' ought to be reserved for the actual movement of particles in a liquid, but it is also often used for this mathematical model of that movement.

Random numbers

The phrase 'random number' refers to one of two ideas. First, as in ideal games with dice or roulette wheels, one number from a finite list is chosen, all of them being equally likely. Second, as in the notion of snapping a stick at a random point, some point in a continuous interval is chosen, no part of that interval being favoured over another. The facility to choose long sequences of such random numbers, each value being independent of all the others, has many applications, as the next section will illustrate.

In 1955, a splendid book One Million Random Digits was published. It follows its title exactly: page after page of the digits zero to nine, grouped in blocks for ease of reading, but successive digits are entirely unpredictable – whatever the recent sequence of digits, you have one chance in ten of guessing the next one. Today, modern computers have built-in software to achieve the same ends. An initial value (the seed) is fed in, and a fixed mathematical formula produces the next value, which acts as a new seed, and so on. There is nothing random about this process at all, and if the same initial seed is used, the same sequence is generated.

But, with a cunning choice of this mathematical formula, the sequence generated passes a battery of statistical tests and looks , to all intents and purposes, as though it were random. The term pseudo-random sequence is used.

No matter how much care is taken in this process, there will always be some lingering fear that hidden flaws in the method used will matter in the use to which the numbers are put. With that caveat, and relying on the experience of a large number of respected scientists, I am prepared to act as though my computer produces acceptable sequences of random numbers on demand. (The obvious danger of insider fraud means that these methods have no place in choosing numbers in Lotteries, or in UK Premium Bonds.)

Monte Carlo methods

How many different numbers will appear on 37 consecutive spins of a standard European roulette wheel? In theory, it could be anything between one and 37, but those extremes would occur very rarely; what is the most likely number of different numbers?

When this problem was first put to me, I did not immediately see an easy way to solve it. There are 37^{37} (a number with 59 decimal digits) possible outcomes of spinning the wheel 37 times, and when you try to write down all the ways in which, say, 28 different numbers could arise, you quickly lose enthusiasm. A more appealing approach was to perform a so-called Monte Carlo simulation .

Here, the computer’s stream of random numbers was used to simulate the outcomes of 37 spins of a wheel, after which the computer counted how many different numbers had arisen. This process was repeated one million times, leading to 24 different numbers on 203,739 occasions, while 23 arose just 199,262 times.

The nearest rivals, 22 or 25 numbers, each happened fewer than 160,000 times. The Law of Large Numbers says that the frequencies of the different outcomes will settle down to their respective probabilities, and these figures essentially settled the matter: the most likely result is that 24 different numbers will arise, and the chance of this is just over 20%.

Days later, I kicked myself for not spotting a standard way to solve the problem! I could calculate the exact probability of getting X different numbers in 37 spins, for any value of X, confirming the conclusion described above. But this does not invalidate the use of simulation to attack this sort of problem - quick and dirty answers can be useful. Indeed, the fact that the simulation gave answers consistent with the exact calculation boosted my general faith that the computer’s random number generator was behaving as intended.

A more serious use of Monte Carlo methods occurs in polymer chemistry. A molecule consists of a large number of atoms, connected along a randomly twisting chain. Atoms can

occur only at places on an evenly spaced lattice and, crucially, no two atoms can be in the same place. How far is it likely to be from one end of the molecule to the other?

We can think of the atoms as being at the places visited by some drunkard, staggering around at random on a three-dimensional lattice for a while, but somehow never visiting the same place twice. Without the requirement that no place be revisited, mathematical experts can make good progress, but that restriction seems to complicate the problem beyond theoretical attack.

However, even a semi-competent computer programmer can write a sensible simulation of this complex, twisting, chain, and, by making one million, ten million, even a billion repetitions, obtain an answer as precise as is required. (Recall de Moivre's work, that precision increases only as the square root of the size of the simulation.)

Or suppose you want to estimate the area of an irregularly shaped leaf. Draw a rectangle around the leaf, and then simulate the positions of a large number of points scattered at random within that rectangle. Your estimate comes from multiplying the area of the whole rectangle by the proportion of points that fall within the leaf's boundaries.

As a final illustrative application, suppose Paul is hoping to set up a new petrol filling station. If he installs four pumps, the minimum viable number, there will be room for up to eight other cars to wait in a queue; each extra pump

removes two waiting spaces, so if he installed the maximum of eight pumps, there would be nowhere to wait. To work out how many pumps will maximize his profits, he can carry out simulations of what would happen if he installed four, five, six, seven, or eight pumps.

As well as the installation costs, the running costs, and the profit margins, he would need to know the rate of arrival of potential customers and the distribution of the time it takes from a car pulling up at a pump until the pump becomes free. He should also take account of the chance that a potential customer would drive past if no pump were free, or the queue too long. All these figures are relatively easy to find or estimate, and it is far cheaper to make computer simulations than to experiment physically with various numbers of pumps over several months. Since he can use the same initial seed each time, he can run all his simulations under precisely the same conditions, thereby improving the comparison between the different estimates of profits.

Why 'Monte Carlo'? Despite the obvious connection between random numbers and casino games, the name was originally just a code word to refer to the use of these methods in military matters, including the early development of nuclear bombs.

Errors in codes

Morse code demonstrates how to transmit messages using only two symbols, 0 and 1 say. But some symbols may get

corrupted, so that an initial 0 arrives as a 1, and vice versa. Even with a low error rate, the message received could mean something quite different from that sent. How best to deal with this?

Suppose each symbol sent has, independently, some small probability of being corrupted. We could repeat the symbols, but a moment's thought shows that sending 00 and 11 rather than 0 and 1 respectively is no use at all: if 01 or 10 arrives, it is a pure guess as to whether 00 or 11 had been sent. We'll guess correctly half the time, but doubling up on the symbols sent means that we should expect twice as many errors, so these factors largely cancel out. But consider transmitting 000 or 111 instead of 0 or 1.

Using 'majority vote' to decode messages, all of {000, 100, 010, 001} will be interpreted as 0, the other four possibilities will be interpreted as 1. If just 1% of symbols sent are corrupted, then when 000 is sent, using the binomial distribution shows that there is a 99.97% chance that one of those four sequences above arrives. That means that the error rate drops from 1% to 0.03%, a factor of over thirty. We would do even better if each digit were repeated five times, but at the expense of the message length. The best choice will depend on the size of the inherent error rate, and the speed of transmission.

Amniocentesis

Prospective parents (and statisticians) Juanjuan Fan and Richard Levine considered whether or not Fan should undergo an amniocentesis - a test to see whether the foetus she was carrying might have Down's Syndrome. Their experience can act as a template for others in a similar situation.

Knowledge of Fan's age, and a simple blood test, put the chance of Down's as 1 in 80; an ultrasound image was encouraging, leading to a use of Bayes' Rule that reduced the chance to 1 in 120. Amniocentesis is an invasive test - a hollow needle is inserted into the abdomen to extract a sample of amniotic fluid; if the extra copy on chromosome 21 that is characteristic of Down's is present, it will certainly be detected, but the test has a risk, estimated in this case at 1 in 200, of causing a miscarriage. Should parents, who would choose an abortion if Down's were present, take this test?

Fan and Levine reached their decision by the logical process of maximizing their expected utility. The worst possible outcome, miscarriage of a foetus without Down's, was assigned utility zero, the best outcome, birth without Down's, was given the utility of unity. To give birth to a Down's child, having opted out of amniocentesis, was allocated utility x, while to take the test and find Down's has a somewhat greater utility, y. (The possibility of a miscarriage is irrelevant in this last case, as the foetus would be aborted anyway.)

The expected utilities with and without the test are made. The test should be taken if the first exceeds the second which, in this case, reduces to requiring that y should exceed (119/200)+ x ; in round figures, y is bigger than 0.6+ x .

If Fan and Levine had felt that the utility of discovering the presence of Down’s, and thus having an abortion, was below 0.6, there would never be any point in taking the test. And the higher the utility they attached to having a child, albeit with Down’s, the higher that threshold would become. If that utility were above 0.4, they should never take the test.

Choosing appropriate values for x and y requires some thought; and if the basic parameters - a 1 in 200 chance of a miscarriage through the test, a 1 in 120 chance of Down’s without the test - were different, the final criterion would change. (See the Appendix.) Plainly, if the chance of Down’s were less than the chance of a miscarriage, it would never be rational to take the test (yes?).

Fan and Levene discussed their dilemma, and their agreed choice of utilities led them to opt for the test. There was a happy ending: no extra chromosome, and no miscarriage.

Haemophilia

Haemophilia is the general name given to a group of disorders that prevent blood from clotting if a cut occurs.

The clotting agent is found on the X chromosome, and the chance it is missing is under 1 in 5,000. As females have two X chromosomes, they would suffer from haemophilia only if it is absent from both copies, giving a chance of under 1 in 25 million, but males possess only one X (and one Y chromosome), so almost all instances of the disease are in males.

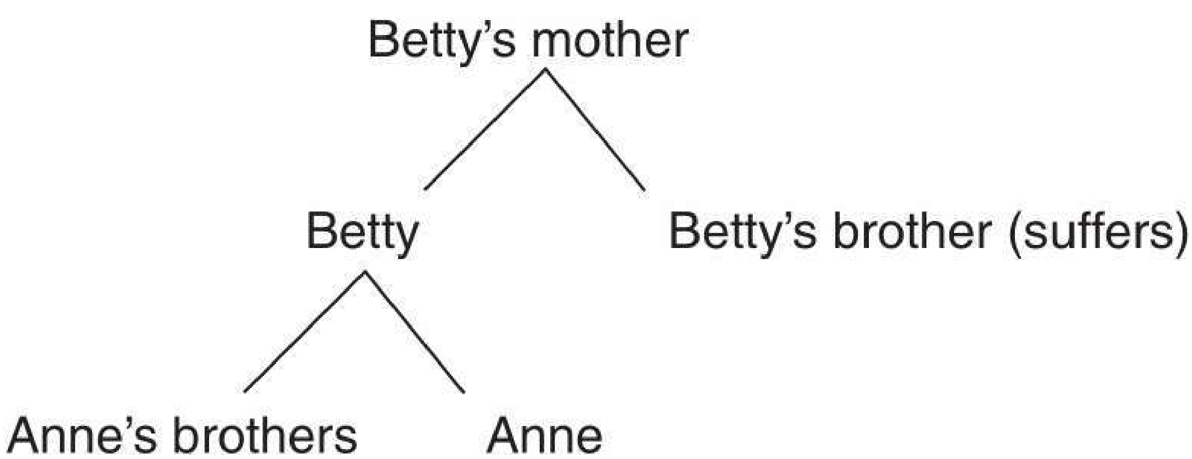
If males do have haemophilia, this becomes known well before they have a chance to father children, but a female may unknowingly have one normal X and one without the clotting agent. Such females are termed carriers, and the chance they pass on the affected gene to any child is 50%. A daughter who receives this affected gene becomes a carrier, a son will suffer from haemophilia. That Queen Victoria was a carrier is certain, as her son Leopold was a haemophiliac, and at least two of her five daughters were carriers. She had three other sons who did not suffer that disease.

Suppose Betty has a brother who suffers from this disease; Betty has several children, including Anne. What is the chance that Anne is a carrier?

To answer this question, it is enough to find the probability that Betty is a carrier; Anne’s chance will always be half that figure. We know that Betty’s mother is a carrier, as Betty’s brother is a haemophiliac. From this information alone, the probability Betty is a carrier is 50%. And if any of Anne’s brothers have the disease, Betty would

be certain to be a carrier, so we look at the case when all Anne’s brothers are healthy.

This situation is tailored for Bayes’ Rule. Let Guilt mean that Betty is a carrier: since the initial probability of Guilt is 50%, the prior odds are unity. If Betty is Innocent (not a carrier), the probability of the Evidence (all brothers healthy) is plainly 100%. But if Betty is Guilty, each brother has, independently, a 50% chance of escaping the faulty gene, so each healthy brother halves the Likelihood Ratio. Turning the posterior odds into probabilities, the chance Betty is a carrier is successively 1/3, 1/5, 1/9, 1/17, . . . , according as she has one, two, three, four, . . . brothers, all healthy.



10. Family relationships

For your own entertainment, suppose that Anne also has sisters, some of whom have sons, and none of these nephews of Anne are sufferers. How does that affect the chance Anne is a

carrier? Check your answer against that given in the Appendix.

Epidemics

The phrase herd immunity expresses the fact that if sufficiently many people are vaccinated, then should a trace of infection enter the community, no epidemic will occur. Thus even those who were not vaccinated are very unlikely to catch the disease. Why should this be so, and how do we discover what ‘sufficiently many’ means?

Typically, those infected may transmit the disease to others, but those who recover have acquired immunity. So we label people as one of Susceptible, Infected, or Removed. The latter are those immune from infection because of vaccination, recovery, physical isolation, or death. It may seem callous to give these four outcomes the same word, but the brutal truth is that, so far as the spread of the epidemic is concerned, they really are equivalent!

To see how an epidemic might develop, let S be the number susceptible and I the number infected. Multiplying these two numbers together gives the total number of possible encounters between an Infected and a Susceptible. Crowded urban populations will be in close contact more often than scattered rural populations of the same size, and the probability of the Susceptible becoming Infected through an encounter will depend on the infectiousness of the disease. Overall, the probability that, during a tiny time period, some

Susceptible individual becomes Infected will be of the form $\beta * S * I$, where β depends on both the infectiousness and how much people mix.

In this same tiny time period, any Infected person may move into the Removed category. Hence the probability that Infecteds reduce by one member will be proportional to the number infected, so takes the form $\gamma * I$, where γ depends on how fast those infected recover, are isolated, or die.

We have found the respective chances of an increase, or a decrease, in the number Infected. The balance between these two probabilities, $\beta * S * I$ and $\gamma * I$, determines whether an epidemic occurs. There is a good analogy with gambling. If the game is loaded against you, each bet is more likely to reduce your capital than increase it: your capital follows a random walk, with an inexorable drift towards zero. But if the game favours you, provided that bad luck does not bankrupt you early, the random walk ushers you far enough from zero to outrun any losing streak. To win a large sum, it is necessary, but not sufficient, for the game to be in your favour.

In the epidemic context, this means that the only time an epidemic (= large fortune) might occur is when any change in the number infected is more likely to be an increase than a decrease. In symbols, $\beta * S * I$ must exceed $\gamma * I$, which is the same as asking that S, the number of Susceptibles, exceeds the ratio γ / β, a quantity termed the threshold for this

population. This is exactly what we were looking for: even if the disease enters into a population,

> an epidemic can only occur when the number susceptible exceeds this threshold.

William Kermack and Anderson McKendrick presented this result in 1927. Epidemics can be avoided by keeping the number susceptible below the threshold, which can be achieved in two distinct ways. First, vaccinate to reduce the number susceptible. Second, find means to increase the threshold. Being a ratio, a threshold increases when the numerator increases - we speed up recovery rates, or isolate infected people faster - or when the denominator decreases - we may be able to reduce the infectiousness, or we can ensure that people mix less, by closing schools temporarily, or postponing sports events where large crowds would gather. We can also assess the likely size of the benefits of these different responses, and so judge which are worth pursuing.

The same principles apply to controlling epidemics in animals. The first step taken to end a foot-and-mouth outbreak in cattle is usually to restrict cattle movements, which increases the threshold size by reducing the denominator. This is often accompanied by mass slaughter (not available with diseases in humans!), which not only reduces the size of the susceptible population, but also increases the numerator in the threshold.

This analysis also explains why we should expect epidemics of childhood diseases at fairly regular intervals. After an epidemic reduces the size of the susceptible population to below the threshold, new births, with insufficient vaccination, gradually take the size above the threshold, so creating the conditions for the next outbreak. The longer the interval between epidemics, the higher the vulnerable population, and the more severe will any epidemic be.

Knowing about probabilities may not cure diseases, but it can help mitigate their effects.

Batch testing

The army wishes to identify which of 1,000 potential recruits may be vulnerable to a certain disease, and thus unfit to serve. A blood test is available, but it costs £50 each time. Can the job be done for less than £50,000?

Provided that only a fairly small proportion will prove vulnerable, the answer is 'yes'. Choose some number K, and pool blood samples from K recruits; then test this pooled sample. If the result is negative, then all those who contributed blood are clear, and need no more testing; otherwise, with a positive result, at least one person in the group would test positive, so we use K more tests, one for each of them, to settle the matter. If we are lucky, one test will suffice, but we might have to make K+1 tests. We hope to make fewer tests than if each person is tested individually.

The best choice of how many samples to pool depends on the probability of a positive test. Suppose this is 1%. Then if we pool ten samples, there will be a positive sample among them about 10% of the time, while the chance all are negative is about 90%. So we will need only one test about 90% of the time, but eleven tests 10% of the time. That leads to about two tests on average. Pooling ten samples reduces the mean cost for those ten recruits from £500 to £100. If we split the 1,000 recruits into 100 groups of size ten, and pool their samples, we expect to save 80% of the initial estimate of £50,000!

More refined calculations show that, when the probability of a positive test is indeed 1%, we would do slightly better pooling groups of eleven, rather than ten, but the difference is very marginal. However, the best choice of the size of a pooling group is quite sensitive to the probability of a positive test. If we expect 2% of recruits to test positive, costs are minimized if we pool eight blood samples; with 5% the best choice is to pool five samples, while if 10% will test positive, pooling four samples turns out best. (Once again, use of the binomial distribution led to these answers.)

In World War II, this simple idea saved the USA about 80% of its initial expected costs.

Airline overbooking

Even though they must compensate passengers who are bumped off fiights for which they have paid, airlines routinely sell more tickets than a plane’ s capacity. Simple economics is the reason: the cost to the airline of making the fiight hardly changes with the number of passengers, but each empty seat is lost revenue. Expecting that not everyone who has booked a particular fiight will show up, how does an airline work out the best amount of overbooking?

Suppose that the plane has one hundred seats, at a fare of £ 100 per seat, but a passenger is paid £ 200 if they have to be turned away because the fiight is full. The airline needs a good estimate of the probability that a passenger who has booked will actually turn up. For charter fiights to holiday destinations, this probability will be close to 100%, but it will be rather lower for passengers who have more fiexibility in their travel plans. Frequency data will help airlines to estimate these chances.

Perhaps each passenger who books has, independently, an 80% chance of turning up for the fiight. If 120 tickets are sold, total revenue is £ 12,000, and, although only 96 passengers will turn up on average, there is a chance, around 15%, that more than 100 show up, and at least one passenger must be left behind. (These numbers again come from using the binomial distribution.) The mean amount to be refunded because of overbooking in this case turns out to be £ 80. Selling five more tickets raises the revenue by £ 500, and the total mean refund increases by only £ 275, so this policy

is expected to be more profitable. The most profitable policy, on average, comes from selling 128 tickets - compared with selling 125, the extra revenue of £300 just outweighs the mean extra refund cost of £295, while 129 tickets would be slightly worse than this.

It is more realistic to suppose that some passengers are more likely to turn up than others, and that groups of passengers who book together will either all turn up, or none of them will. But these details can be incorporated into the model, and airlines will continue to sell seats until the expected cost of compensation exceeds the extra revenue.

Queues

One of the best-developed applications of probability is the study of queues of various kinds. The initial impetus came from attempts to understand congestion in telephone lines - the work of the Danish telephone engineer Agner Erlang is honoured by the use of his name as the unit of the volume of telephone traffic. Queuing theory contributed to the success of the Berlin airlift of 1948/9, and the systematic study of queues flourished during the next twenty years.

David Kendall introduced notation, with the format A/B/ n , now universally accepted as a shorthand way of describing queues where customers arrive singly. The first component, A, refers to the distribution of the time between customer

arrivals, while B describes the distribution of the time it takes to serve a customer, and n is the number of servers.

For example, in the expression D/D/3, 'D' is short for deterministic, meaning that there is no randomness at all. Customers arrive on the dot at fixed time intervals, all service times also have exactly the same length, and there are three servers. This queue would be of little interest in the field of probability, as there is no variability. But suppose there is a huge number of potential customers, each of whom has a tiny chance of turning up in a given short time interval, so that customers arrive at some overall average rate, but completely at random. Here the symbol M is used, to honour Andrey Markov, so M/D/2 means that customers arrive at random, and select either of two servers, who each take some fixed time to do their job.

We want to know how queues will behave. The main questions are how long do customers have to wait, how frequently are servers just sitting on their thumbs, and what can we do to improve matters? The 'servers' may be the intensive care beds, while 'customers' are patients needing that care.

If customers arrive every five minutes on average, and there are three servers, then unless the mean service time is less than fifteen minutes, a queue of indefinite length will build up, and the whole operation is unsustainable. So we must assume that the mean service time, taking account of the

number of servers, is less than the mean time between arrivals. The ratio of these two mean times is termed the traffic intensity , some figure between zero and unity.

In an ideal world, customers would face only short queues and the server would be busy nearly all the time. But these two requirements are diametrically opposed. Take a simple case with one server and customers arriving at random. If the traffic intensity were 0.9, calculations show that we might expect about five waiting customers on average, and an empty queue about 10%

of the time. If the intensity rose to 0.98, the server would be unoccupied just 2% of the time, but the mean queue length would shoot up to 25. Most customers would consider this a worse arrangement. Unless servers have enough 'idle time', customers will get angry, leave the system, or do both.

The queue behaviour depends on much more than just the traffic intensity. Other things being equal, the more variable the service time, the longer you can expect the queue to be. With several servers, it matters whether they operate as in my railway station, where one central queue feeds up to six servers, or as in my supermarket, where I choose which aisle to join, and stay there. In some situations, e.g. ambulance calls, some customers may have higher priority. Many queues use the 'First come, first served' rule, but when non-perishable goods are stored on a

shelf awaiting use, ‘Last come, first served’ may apply. Some queues feed other queues, the servers may work at different speeds, bunches of customers could arrive together. Eagle-eyed queuing theorists have found answers to the central questions under almost any realistic model you can think of.

Chapter 8 Other applications

The applications of probability away from dice, casino gambling, and various aspects of natural science can be overlooked. In this chapter, I have picked out some of its appearances in law, social science, sport, and economics to emphasize its ubiquity. The common theme is that decisions we make will depend on the probabilities of various outcomes, so we need methods that lead to reasonably reliable estimates of those different probabilities.

Legal matters

Although Lord Denning, one of the best-known UK judges in the 20th century, had a mathematics degree, few lawyers feel comfortable with probability. This ought to be astonishing, as phrases relating to the subject are used freely in courts. In civil cases, such as libel, to say ‘on the balance of probabilities’ clearly puts the dividing line at 50%. But in criminal cases, where a jury is asked to convict only if they are ‘sure’ of Guilt, there is no consensus on a figure. Some people would wish to convict if they were 80% certain of Guilt, others would use 95% or even higher. These are plainly subjective probabilities. And although the same phrase is used whatever the offence, some would apply a lower threshold of proof for a relatively minor offence. This could make it harder to convict mass murderers than fare dodgers.

Suppose an expert witness testifies that the DNA of the accused matches DNA found at a crime scene, and that the chance of a match between the latter and an innocent person chosen at random is one in several million. Jurors may have two distinct problems with this statement. The first is that they may think that it is equivalent to saying that the chance the crime scene DNA is NOT that of the defendant is one in several million. The second is that they may treat all such tiny figures as equivalent, even though one in ten million differs from one in a billion by a factor of a hundred.

The first error has been termed 'The Prosecutor's Fallacy'. Starkly, it is equating the chance of Innocence, given a DNA match, to the chance of a DNA match, given Innocence. This is logical nonsense: the chance of zero arising, given a fair roulette wheel, is not the same as the chance that the wheel is fair, given that zero occurred. This trap can be avoided by giving the jury an estimate of how many citizens might match the crime scene DNA. With a population of around 60 million, if the match chance is one in 2 million, there might be thirty or so; if it is 1 in 20 million, there might be about three; it is unlikely there are more than half a dozen. But do not overlook the phrase 'selected at random': the more close relatives the criminal has, the more matches we would expect, the less strong this evidence against the accused.

The second error can best be avoided by remembering how Bayes' Rule measures the usefulness of any piece of evidence. Before this evidence is presented, you have some idea of the odds that the accused is Guilty. If the evidence is ten times more likely under Guilt than Innocence, then the odds of Guilt get multiplied by ten: while evidence that is three times more likely under Innocence than Guilt reduces the odds by a factor of three, and so on. With DNA evidence, it often happens that the chance of the evidence, assuming Guilt, is 100%, which makes the impact of the evidence clear: the odds of Guilt should be multiplied by whatever the 'several million' figure actually is.

Randomized response

A head teacher wishes to ascertain what proportion of his senior students smoke cannabis. Direct questions are unlikely to produce truthful answers, but a technique known as randomized response is available. The main idea is that the teacher recording the answers does not know what question is actually being asked, so that cannabis users are able to answer honestly without fear of being identified.

The words 'I smoke cannabis' are written on each of 80 cards, and 'I do not smoke cannabis' on another 20. Each card is placed in an identical envelope, these 100 envelopes are mixed thoroughly in a large bag. The students should see this operation being done, so that they know that the bag

contains both versions of the question, and in those proportions.

Angela selects one envelope at random, opens it, reads the question to herself, and simply says either ‘Agree’ or ‘Disagree’. She then puts the postcard back in the envelope, returns the envelope to the bag, and shakes the bag up ready for the next student.

Suppose that one-third of the responses are ‘Agree’. Because the students are picking the envelopes at random, ‘Agree’ is the honest response from 80% of users, and 20% of non-users. A few lines of algebra show that this is consistent with 2/9 of the students being users. The head teacher has his answer, no individual student has been identified.

Alternatively, replace ‘I do not smoke cannabis’ with a question on an unrelated subject, for which the proportion of ‘Agree’ answers is known. If a previous survey has established that half the students own a pet, and there is no reason to link pet-owning with cannabis smoking, the statement on 20 cards could be ‘I own a pet’. Then, if one-third of the responses are ‘Agree’, we estimate that 7/24 of the students are users.

The calculations giving these estimates are in the Appendix.

The uncertainty as to which question is being asked each time leads to some imprecision in the final estimate. The proportion of envelopes that contain the sensitive question should be as high as possible, but low enough for genuine cannabis users to believe that giving an honest answer will not have repercussions. Putting the sensitive question on as many as 95% of the postcards would not work.

WADA

The World Anti-Doping Agency seeks to promote sport as a healthy activity, by identifying athletes who take performanceenhancing drugs, and excluding them from competitions. But whatever methods are used, any testing programme is liable to two opposing types of error: claiming that an athlete is using drugs when they are innocent, and passing an athlete as clean when they are a drug user.

Unfortunately, methods of reducing the chance of making either type of error often tend to increase the chance of the other. For example, one test measures the ratio of testosterone to epitestosterone. The body normally produces these substances in fairly equal amounts, but those athletes who seek to cheat by injecting testosterone will have a high T/E ratio. Athletes whose ratio is above some specified amount, say six to one, will be banned. However, the T/E ratio varies naturally: it changes over a menstrual cycle, it will increase if you catch flu. Set the critical T/E ratio

too high and no drug cheat will fail it; set it too low, and many innocent athletes will be wrongly accused.

Suppose the chance that a particular test makes a mistake is 1%. That means that if the athlete is innocent, the chance they fail is 1%, if they are users the chance they pass is also 1%. Sam fails the test: what is the chance she is innocent?

Put like that, the temptation to say '1%' looks overwhelming – this test gets things wrong one time in a hundred, so if it says she has failed, that will be wrong one time in a hundred. Resist this temptation. The only valid answer is 'We do not know. It could be any figure. We need to know the proportion of drug cheats in the population.'

For, suppose that proportion is 1% or so. Then among 10,000 athletes we expect 100 drug cheats, and 9,900 innocents. In testing, we expect just one drug cheat to pass, leaving 99 who fail. But 1% of the 9,900 innocents, i.e. another 99 athletes, will also fail the test. Among those who fail, half are innocent: the chance Sam is innocent would be 50%.

If the proportion of cheats differs from 1%, this conclusion changes. If it is higher, the chance that Sam is innocent will be less, but if it is lower, her chance of innocence will be even higher. The lower the proportion of drug cheats, the less satisfactory is this test, despite its apparently impressive performance.

This same logic applies when we consider how to detect potential terrorists at airports. Whatever screening devices are used, they cannot be perfect, but suppose that the probability a real terrorist evades these checks is tiny, 1/10,000, while the chance that an innocent person is led away for intensive interrogation is a minuscule 1/100,000. How likely is it that someone picked out is guilty?

We cannot answer the question without having some idea of the proportion of would-be passengers who are terrorists. Try one in a million - frighteningly high, given that Heathrow handles over fifty million passengers a year. But the figures assure us that, even with fifty potential terrorists, it is overwhelmingly likely that all will be detected.

Unfortunately, five hundred innocent passengers will also be detained! Among those stopped by this system, fewer than 10% are terrorists. And if there are fewer than fifty terrorists, the chance that someone who is stopped is indeed guilty is even lower. Detection methods must have much better performance figures if they are to be useful.

Football results (1)

Betting on the results of soccer matches generates a substantial interest in the UK. All sorts of exotic bets can be made - on the time of the first throw-in, the sum of the shirt numbers worn by all the goal-scorers, how many red and yellow cards will be ? ourished during the game - but most interest is on which of the three results, Home win, Draw, or

Away win, will occur. A rational punter will assess the respective probabilities of these results, and his decision on whether to bet, and how much, will rest on these assessments as well as the payout prices offered by the bookies.

But how might the punter deduce his degrees of belief in the different outcomes? In May 2009, statistician David Spiegelhalter took up the challenge on BBC Radio's *More or Less* by analysing the ten games to be played in the Premier League two days later. For each game, he estimated the number of goals each team might score, on average, taking account its own strength in attack, and their opponents' defensive capabilities. For example, a strong Home team (Arsenal) were estimated to score 2.1 goals, on average, against Stoke City.

No team can score 2.1 goals, but that figure is just the qaverage over a hypothetical number of matches. The crucial step is to assess the probabilities of 0, 1, 2, 3, . . . goals in a single game, and Spiegelhalter used the Poisson distribution. Data over many years show that this is pretty good at describing how the actual number of goals tends to vary around its average. With Arsenal's figure of 2.1, the chance of no goals came out as 12%, one goal as 26%, two goals as 27%, three as 19%, and so on.

Data for Stoke put their mean score as 0.67 goals. This translates into a 51% chance of no goals, a 34% chance of just one goal, 11% for two goals, and so on. With a leap of

faith, take the numbers of goals scored by each team as independent. So the probability of a 2-1 score comes from multiplying the chance the Home team scores twice by the chance the Away team scores once - in this case, 27% * 34%, around 9%.

In this way, the probability of any possible score is estimated. Then the probabilities for each of Home win, Draw, and Away win are found from the Addition Law, by adding up the separate probabilities of all the scores that lead to those three respective results. This gave Arsenal a 72% chance of victory, Stoke had a 10% chance, leaving a chance of 18% for the Draw. The score given the highest probability, at 14%, was 2-0.

Do not scoff! In the ten games, the exact score given the highest probability happened twice, and eight of the ten match results were those that were identified as being the most likely. A betting man, who had placed money on each 'most likely result' and on each 'predicted' exact score, would have smiled happily as the match scores unfolded.

How can we reconcile a 72% degree of belief that Arsenal would win that match with ideas of frequency, as there is no question of playing this game hundreds of times and counting how often Arsenal won? Recall how we judged the reliability of a weather forecaster when she says that the chance of rain tomorrow is 30%:

there is only one tomorrow, it will either rain or it will not. However, we can look at all the occasions when she gives rain a 30% chance, and check its actual frequency. We shall believe her claim about tomorrow, or not, on the basis of her overall record. With soccer matches, we can make similar calculations for all games played over the season. Among these, there might be forty or so where some result was given a probability close to 72% - we can check whether the 'predicted' result did occur with a frequency around 72%, as a way of validating our methods.

Can a gambler expect to make money by using these ideas? The payout prices depend heavily on how much is staked on each outcome, and the largest sums are usually staked on one team or the other to win. Bets on a Draw tend not to attract committed fans. If the chance of a Draw is assessed as 25%, and the payout price is better than three to one, the opportunity to profit is there.

Do not assume that the best bet is on the outcome with the highest predicted probability!

Football results (2)

Before the 2010 soccer World Cup Finals began, statistician Ian McHale published the results of his calculations, which allocated to each of the 32 teams some non-zero probability of winning the trophy. He made Spain the favourites, albeit with a winning chance of only 11.6%, followed by Brazil, whose chance was put at 10.3%.

To obtain these figures, McHale used an approach similar to that described above for each match. However, he did not make a direct calculation of the probabilities of the distinct match outcomes, he relied on a Monte Carlo simulation.

Thus, for a match in which England’s mean score was put at 1.5 goals, the Poisson model gives a 22% chance of no goals, a 33% chance of one goal, and so on. The computer’s random number generator selected one of the values 0, 1, 2, 3, . . . with the appropriate probabilities, and did the same thing for England’s opponents, leading to some simulated score such as a 2-2 draw. Similar simulations were made for every scheduled match, leading to simulated group tables, and then to matches in the knockout stages all the way to the final. This process was repeated 100,000 times, and the number of simulations in which each team emerged as champions was recorded. Spain ‘won’ 11,633 times, hence the 11.6% figure noted earlier. The Law of Large Numbers, as usual, is the justification.

And Spain did win! Were McHale’s probabilities ‘correct’? We cannot know. Perhaps Spain would have won 65% of the time, had it been possible to make indefinite repetitions of the tournament. But the best evidence that his methods make good sense is that bookmakers follow a similar path to set their initial payout prices to attract punters.

Black-Scholes

Share prices on stock markets fluctuate, sometimes for no apparent reason. If the price is £5 today, you do not know what the price will be next month. However, you can buy an option - the right to buy (or sell) that share at the strike price of £5.20 at a given future time. If, at that time, the market price is less than £5.20, you will not exercise your option to buy, but if it is above that price, you can make an instant profit by taking up the option,

and immediately selling. Corresponding remarks apply to an option to sell. What are fair prices for these options?

Fischer Black and Myron Scholes addressed this question in 1973. At the heart of their work was the assumption that the changes in share prices varied randomly, but in a particular way related to the Gaussian distribution. The fair prices for both buy and sell options were found to depend on the current price, the price at which the option would be exercised, the intervening time period, prevailing interest rates, and the volatility of the underlying share price (as measured by the standard deviation over a period): but not on the mean amount by which the share price was expected to change!

This last point may be surprising, but that is how things work out. It is also quite useful, as it means that we have no need to add to any uncertainty by estimating the trend in prices. If you want to discover the fair price for some particular option, free software is widely available - just

type 'Black-Scholes' into your favourite search engine. Given the current price and the strike price, the fair cost of a buy option would increase if the time period were longer, or if interest rates were higher, or if the volatility of the share price were higher.

How do the claims in this last sentence accord with your intuition? The first does seem reasonable, as the longer you are prepared to wait, the higher the chance of an increase in the underlying share price, but the other two claims are more subtle. An increase in volatility also increases the chance of a jump in the share price, but it also happens to decrease the mean change in price - and the former effect turns out to be bigger.

The volatility is measured by looking at the changes in the share price over the 250 or so trading days in one year. This should give enough data for an estimate to be reliable, but not stretch so far into the past as to be irrelevant for current conditions. A poor estimate of the volatility will lead to an unreasonable price for an option.

A model is only useful when its key assumptions are not violated. And, as Figure 7 shows, taking the Gaussian distribution as a model for price fluctuations implies a really tiny probability for catastrophic events, such as the price dropping by more than three or four standard deviations. When the actual probability of such an event is significantly underestimated, the model is undermined, and

the conclusions it indicates may have no sound basis at all. The extreme-value distributions, mentioned in Chapter 4, have been used to address this problem.

Share portfolios

Companies A and B are both expected to make profits. With low interest rates, A is expected to return 20%, while B should return 40%; with high interest rates, the positions are reversed – A should gain 40%, B 20%. Suppose Nick is a risk-averse investor, while Mary is risk-attracted.

If low or high interest rates are seen as equally likely, both companies may look equally attractive, with a mean return of 30%. In accordance with their respective attitudes to risk, Nick could divide his funds equally between the two companies, and guarantee to get 30%, whether rates are high or low: Mary could plump for one company or the other, hoping to get 40% but accepting she might get only 20%.

Suppose B is replaced by company C, which will return 10% with low interest rates, or 50% if rates are high – again an average of 30%, like company B. But now mixing A and C makes no sense to either investor: Nick prefers A alone, Mary puts everything into C.

The essential difference is that the returns from A and B are negatively correlated – in conditions when one is high, the other tends to be low; but returns for A and C are positively correlated – they do better or worse together. ‘Correlation’

is measured on a scale from - 1 (total negative correlation) to +1 (total positive correlation). If two assets fluctuate in value independently of each other, their correlation will be zero.

Risk-averse investors are encouraged to diversify their holdings, so that any losses might be balanced by gains elsewhere. They wish to hold negatively correlated assets. But there is an inescapable piece of logic: if X is negatively correlated with Y, and Y is negatively correlated with Z, then X and Z will tend to be positively correlated!

However, all is not lost. A mathematical result, due to Salomon Bochner, proves that it is indeed possible for each pair of assets in a large portfolio to be negatively correlated; but the greater the number of assets, the harder it is to achieve mutual negative correlation.

Chapter 9 Curiosities and dilemmas

At the beginning of this book, I noted that some aspects of probability appear, at first sight, to defy common sense. Examples have arisen as the story has unfolded. Here are some other circumstances where intuition can be misleading, but, with sufficient care, these apparent contradictions can be explained. The subject of probability is wholly free from real paradoxes. But although ideas of probability can help us make sensible decisions, we may also find that even thinking about the probabilities of certain events might lead to uncomfortable dilemmas.

Parrondo's paradox

Graham Greene's novel *Loser Takes All* is a fine read, but based on a false premise: that there is some clever mathematical way of combining bets on a roulette wheel to give the player an advantage, rather than the house. On the contrary: mathematics has proved that, when all individual bets favour the house, no combination whatsoever can turn matters round and favour the player. Sorry, folks.

Juan Parrondo has shown that you have to be very precise in how you formulate a general claim that, whenever all bets favour one side, it is impossible to combine bets so that the other side has an advantage. I describe here a modification of his idea, due to Dean Astumian, who described a simple game played on the board with five slots, shown in Figure

11 . (This is not a serious game. It was constructed merely to make this point.)



11. The board for Astumian’ s game

You need some way of generating a random event that will occur 1% of the time: perhaps a bag with 99 White balls and one Black ball, or a spinner that is equally likely to come to rest on any one of its one hundred sides. To begin the game, place a token on the slot marked ‘Start’ . Every move will take the token one step left, or one step right, and you win if the token reaches Win before it hits Lose.

There are two basic sets of rules, call them Andy and Bert. With Andy, from Start you always move to Left, and from Right, you always move to Win. From Left, you use the spinner, to give a 1% chance of moving to Lose, and a 99% chance of moving back to Start. With Bert, the spinner is used for Start to give a 99% chance of moving to Right, and a 1% chance of going to Left. From Right,

you always return to Start, while from Left, it is the same as in Andy - the spinner gives a 1% chance of moving to Lose, a 99% chance of returning to Start.

Analysis of these games is simple. In Andy, there is no provision ever to reach Right; you shuffle around between Start and Left, until random chance takes you from Left to Lose. In Bert, you usually shuffle between Start and Right, with occasional visits to Left. Eventually, on one of these sojourns to Left, random chance takes you to Lose. In either game, the chance of reaching Win is zero.

For the new game, Chris, you also need a fair coin. At each turn, toss this coin: if it shows Heads, use the rules of Andy, if it shows Tails, use the rules of Bert.

It turns out that your winning chance in Chris exceeds 98%! It is easy to see why you are strong favourite: if ever you get to Left, you are overwhelmingly likely to return to the safety of Start. From Start, you play Bert half the time, with its 99% chance of getting to Right; and in Right, you play Andy half the time, inevitably winning.

Following either Andy or Bert, you must lose: flip between these games at random, and you win nearly every time! Framing a mathematical theorem that excludes examples like this, but confirms that Greene’s plot rests on shaky ground, requires very precise language!

2+2=4, or 2+2=6?

Suppose we carry out Bernoulli trials with a fair coin, i.e. each toss, independently, is equally likely to be Heads or Tails. A typical outcome will be HHTHTTTHT. . . . The mean

number of tosses until Heads appears is two; but what is the mean number of tosses until we see HT, or HH?

The intuitive answer is four, as we expect to wait two throws for the first symbol, then another two throws for the second. And the mean number of throws until we see HT is indeed four, but this is not the case for HH. To see that pattern, the mean number of throws is six!

The reason for the difference is that, to get HT, it is correct to argue that we expect to take two throws to get the H, then another two to get the T that completes the pattern. And Two plus Two equals Four. But for HH, after we have the first H, the next throw will be T half the time, and we must begin again - all throws up to that point will have been wasted. The algebra leading to the correct answer is in the Appendix.

Between H and T, each is equally likely to appear first; what about between HH and HT? Again, each is equally likely to arise before the other, since we must wait for the first Head, and then the next throw determines the answer. However, between HH and TH, the latter is three times as likely to appear first! The reason is simple: the sequence will begin with HH one-quarter of the time, but unless this happens, it is inevitable that TH appears first. (Think about it.)

The game Penney-ante is based on the above ideas. You invite your opponent to select any of the eight possible triples like HHT, or THT, etc., that might occur in three

consecutive throws of a fair coin. You select a different one, a neutral person tosses the coin repeatedly, and the winner is the person whose triple is seen first. Despite the apparent generosity of allowing your opponent to have first pick, this game favours you - if you know what you are doing. Whatever she chooses, you can select a triple that will appear before hers at least 2/3 of the time! The winning recipe is in the Appendix.

Give me a clue…

(1) Three double-sided cards of identical shape and size are placed in a bag. One card is Blue on both sides, one is Pink on both sides, the last is Pink on one side, Blue on the other. One card is selected at random, and one side of it is exposed, and seen to be Pink. Is the other side more likely to be Pink or Blue? Or are the chances equal? Over to you - answer below.

(2) Careful counting shows that a bridge hand of thirteen cards, dealt from a well-shuffled deck, will contain two or more Aces about 26% of the time. You deal out a hand to Lucy. To the question ‘Does your hand contain at least one Ace?’, she answers ‘Yes’. On a separate occasion, you deal a hand to Tina, and ask ‘Does your hand contain the Ace of Spades?’ She also responds ‘Yes’. Which of the two hands is the more likely to contain two or more Aces? Or are the chances equal? See below.

(3) Suppose that, among 1,000 males and 1,000 females, all with satisfactory qualifications, 480 males but only 240 females gain admission to a university. A clear case of sex discrimination – men are twice as likely as women to be admitted?

The answers? With the Pink/Blue cards, seeing a Pink side plainly eliminates the double Blue card. All three cards were equally likely, just two are left, Pink/Pink or Pink/Blue. With one of these cards, the reverse side is Pink, but with the other card, the reverse side is Blue. It looks as though Pink and Blue are equally likely.

That reasoning is sloppy: Pink is twice as likely as Blue, and you can check this by doing this experiment a dozen or so times. Better, note that the cards have three Pink sides between them, and all of those are equally likely to be the side seen. But only one Pink side has Blue on the reverse – twice, a Pink side also has Pink on the reverse. (You could use Bayes' Rule, but that would be sledgehammer and nut territory.)

As an impoverished graduate student, Warren Weaver, one of the founders of Information Theory, taught other students the usefulness of understanding probability by consistently winning money from them when playing this game.

With the deck of cards, we know both times that the hand has at least one Ace, and many people would suggest that Tina and Lucy are equally likely to hold two or more Aces – all

Aces are equally likely, so why should Tina confessing to the Spade Ace in particular make any difference? Reject those thoughts, and do the counting properly.

For Lucy, among the hands with at least one Ace, we find the proportion that have two or more - it is about 37%. For Tina, along with the Ace of Spades, she has twelve more cards, chosen at random from the remaining fifty-one. About 56% of the time, these include another Ace: Tina is far more likely than Lucy to have two or more Aces.

Your suspicious mind rightly tells you that the answer to the third question is ‘No’. For suppose that, in the English department, 20% of 950 women who applied, and 10% of the 50 men, were admitted; in Business Studies, all 50 women gained admission, but only half the 950 men. Do the sums: 240 women and 480 men were successful, but, in each department, the success rate for women was twice that for men. Any discrimination was against men, not women!

Indeed, in real life, from thousands of applications to Berkeley’s Graduate School, 44% of males but only 35% of females were admitted. However, when the applications were broken down into the separate departments, there was hardly any difference between the admission rates of men and women. But admission rates did vary between departments, and female applications were largest to those departments that admitted a smaller proportion from both sexes.

This counter-intuitive result is an example of Simpson’s Paradox . It shows the perils of working with proportions, rather than absolute numbers, and it turns up all over the place.

It is far more than a mere curiosity. You have no justifiable claim to be numerate unless you know what it is about.

Do you want to know?

I have argued that probability is the key to making decisions under uncertainty, and I will not retreat from that. But the ability to know probabilities more precisely, and in new circumstances, throws up some uncomfortable dilemmas.

It is now possible for individuals to have their own entire genetic code sequenced, but Nobel Laureate James Watson and Harvard psychologist Steven Pinker have both opted not to know which version of a gene known as APOE they carry. Having one copy of the epsilon4 version of this gene quadruples the chance of developing Alzheimer’s disease, while having two copies is associated with a twenty-fold increase in the chance. (Paradoxically, having this epsilon4 version is also associated with benefits during one’s youth.) Another Nobel Laureate, Craig Ventner, knows that he does have one copy of epsilon4. One research laboratory has the policy of never disclosing to volunteers their APOE

status, on the grounds that, with current knowledge, there is no treatment available to mitigate bad effects.

But some commercial companies may be very interested in your APOE status, indeed in your whole genome. If your genetic composition suggests a high probability of an early death, they may be willing to give greatly enhanced annuity rates - but may also demand much higher medical premiums. Companies who have full genetic information on an individual might 'offer' a bespoke service, tailored exactly to the life prospects of the client.

John and Tom are both 65 years old, and will each use £15,000 to buy an annuity; normal life expectancy is, say, 15 years, but John's genes suggest 10 years longer, Tom's 10 years less. Ignoring genetics, Company A offers both the same sum, £1,000 per year. But company B uses the genetic information, and offers Tom £3,000 per year, but John only £600 per year. Recall the aphorism that, in the long run, averages rule. Both men would accept their higher offer, so company A should expect to pay out £25,000 to men like John, a loss of £10,000 each time, while company B expects to pay £15,000 to Tom and his ilk, and so break even. Company A will collapse, B will survive.

If the only viable insurance companies are those like B, we can expect many miserable people, who either cannot get medical/travel insurance at all, or who discover that their

retirement plans are thrown askew because they are forecast to outlive their savings.

Barristers are advised to ask, in cross-examination, only those questions to which they already know the answers. Before you ask for your genome to be sequenced, assure yourself that you are fully prepared for what you might learn. Think of all the stages in life: a printout of your child's genome at birth may have devastating news; when contemplating marriage, should you and your betrothed take steps to learn the likelihood of any children being severely afflicted? Should your employer have the right to deny you promotion because you have an increased risk of some illness? Should candidates for high public office, say president or prime minister, have to disclose their genome, so that voters are more aware of any genetic propensity to become unstable?

A randomly selected UK female has a 12% chance of contracting breast cancer. But if she has inherited certain mutations of genes known as BRCA1 or BRCA2, the chance increases to 60%. Should a mother of three daughters, whose sister has this mutation, be tested herself? And if she is tested and gets bad news, at what ages (if ever) should her daughters be told that they each have a 50% chance of having inherited this mutation?

Whatever you feel in such uncomfortable circumstances, recall that it is 'probability', and not 'certainty'. If the probability of having the mutation is 10% for Emma, and

60% for Fiona, it may well turn out that Emma develops breast cancer, while Fiona does not. If they know their chances of having this mutation, they have to deal with this knowledge in their own way. To repeat the central dogma of decision theory: the rational action is the one that maximizes the mean utility of the consequences. You can never be sure you have taken the action that would have worked out best, but you have made optimal use of the information you have. You cannot ask for more.

Appendix

Answers to questions posed

Chapter 2: With fair dice having eight or ten sides, the chance of an even number is 1/2, and the chance of a multiple of six is 1/8 or 1/10 respectively. The respective chances of a multiple of three are 1/4 (two in eight), which gives independence, and 3/10, which does not.

Chapter 6: In the Colour of Money puzzle, you should aim to bank £12,000 at your first choice. If it works, you hope to bank £3,000 next time; if it fails, you must try for £15,000 next time. Your overall winning chance is $(6/12)*(10/11)+(6/12)*(4/11)=84/132=7/11$, which is higher than any alternative tactic.

In 46 similar poker games, the nine times a Spade arises, you make a net profit of 50 chips; the other 37 times, you lose the extra stake. If you can stay in for 12 chips or fewer, you'll be ahead on average, but if you have to pay 13 chips or more, you'll expect to lose more than you gain.

Chapter 7: The chance of a miscarriage is 1 in m, the chance of a Down's baby is 1 in n, both m and n large and $m>n$. Take the test if $y>x+n/m$.

Suppose Anne's sister Celia has n sons, all healthy. If Betty is not a carrier, all Celia's sons are sure to be

healthy; if Betty is a carrier, the chance Celia will not be a carrier (and hence all her sons are healthy) is 1/2, and the chance she is a carrier, but nevertheless all her sons are healthy, is $(1/2)^{n+1}$. So the posterior odds of Betty being a carrier, after the information about all Anne's healthy nephews, are the prior odds (i.e. those computed using the information about Anne's healthy brothers), multiplied by all these factors of the form $(1/2+(1/2)^{n+1})$ from her sisters. Convert this figure into the probability that Betty is a carrier, and halve it to find the chance Anne is a carrier.

Chapter 8: In the randomized response examples, suppose the proportion of smokers is x, and all answers are truthful. Then, in the first case, the overall proportion of 'Agree' answers is $0.8x+0.2(1-x)$; equate this to 1/3, solve to find $x=2/9$. In the second case, the proportion of 'Yes' answers is $0.8x+0.2/2$; setting this equal to 1/3 gives $x=7/24$.

Chapter 9: Let x be the mean number of throws to get HH. To get the first H takes two throws, on average. After this, we need at least one extra throw; and even then, half the time we must begin again. So $x = 2+1+(x/2)$, hence $x = 6$.

In Penney-ante, if your opponent selects HHH, you should choose THH, and your winning chance is 7/8; if she picks HHT, again select THH, and win 3/4 of the time; to her HTH,

respond HHT, against THH, use TTH - in both these cases, expect to win 2/3 of the time. Symmetry leads to the best choices against TTT, TTH, THT, and HTT.

References and further reading

Chapter 1

I. J. Good, *Probability and the Weighing of Evidence* (Griffin, 1950). Jack Good worked with Alan Turing on the Enigma project at Bletchley Park, and later at Manchester University.

L. J. Savage, *The Foundations of Statistics* (Wiley, 1954). Jimmie Savage was a leader of the subjective approach to probability.

Chapter 3

W. Feller, *An Introduction to Probability Theory and Its Applications* , Vol. I (Editions, 1950, 1957, 1968). Unsurpassed in its influence.

I. Hacking, *The Emergence of Probability* (Cambridge University Press, 1975). Authoritative, highly respected.

A. N. Kolmogorov, *Foundations of the Theory of Probability* , 2nd edn. (Chelsea, 1956). (The original publication in 1933 was *Grundbegriffe der Wahrscheinlichtkeitsrechnung* .)

S. M. Stigler, *The History of Statistics* (Harvard University Press, 1986). Scholarly work, meticulously researched.

I. Todhunter, *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace* (Chelsea, 1949). The chapter on Laplace contributes almost a quarter of this comprehensive work.

Chapter 4

D. J. Hand, *Statistics: A Very Short Introduction* (Oxford University Press, 2008). Very readable account from a President of the Royal Statistical Society.

Chapter 5

B. Goldacre, *Bad Science* (Harper Perennial, 2008). Mainly about misuses of statistics in medical science, full of good sense.

Chapter 6

N. Henze and H. Riedwyl, *How To Win More* (A. K. Peters, 1998). The subtitle 'Strategies for increasing a lottery win' tells it all.

E. O. Thorp, *Beat the Dealer* (Vintage Books, 1966). Made Thorp famous, but his fortune came from the stock market, not casinos.

Chapter 7

J. Fan and R. A. Levine, 'To Amnio or Not To Amnio: That Is the Decision for Bayes', *Chance*, 20(3) (2007). The full

account, outlined in this book.

RAND Corporation, *One Million Random Digits, with 100,000 Normal Deviates* (RAND, 1955). Contains exactly what it says on the tin.

Chapter 8

Significance, 2(1) (2005). This journal contains several useful articles about probability and the law, including Peter Donnelly's 'Appealing Statistics', and Tony Gardner-Medwin's 'What probability should a jury address?'.

Websites and other publications

Several public-spirited groups and individuals post or update material germane to probability on the internet. In no particular order, I recommend:

http://www.dartmouth.edu/~chance/ takes you to the Chance website of Dartmouth College, with a variety of useful links, including the archive of 'Chance News', video and audio material.

http://www.mathcs.carleton.edu/probweb/ is aimed at researchers and teachers, but others will also be pleased to see the list of books, newsgroups, pertinent quotes, and miscellaneous information.

http://www.plus.maths.org/ The very readable online Plus magazine frequently contains probability material among its

articles.

For accurate information on gambling odds in a host of games, Michael Shackleford has set up http://www.wizardofodds.com/ where your query may well be answered.

Significance is the magazine of the Royal Statistical Society. No formal qualifications are needed to join the RSS, merely an interest in statistics.

Statistical Science is a mainstream academic journal; several issues contain highly illuminating 'conversations' with distinguished probabilists who talk candidly about their careers.